Semantic Processing of Legal Texts (SPLeT-2012)

Workshop Programme

9:00-11:30 – General Session

9:00 –09:10 – Introduction by Workshop Chairs

9:10 –09:40
Giulia Venturi, *Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts*

9:40 –10:10
Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, *Using Legal Ontology to Improve Classification in the Eunomos Legal Document and Knowledge Management System*

10:10–10:30
Antonio Lazari, Mª Ángeles Zarco-Tejada, *JurWordNet and FrameNet Approaches to Meaning Representation: a Legal Case Study*

10:30 – 11:00 Coffee break

11:00–11:30
Lorenzo Bacci, Enrico Francesconi, Maria Teresa Sagri, *A Rule-based Parsing Approach for Detecting Case Law References in Italian Court Decisions*

11:30-12:30 – Session on the Results of the "Collaborative Annotation Exercise"

11:30–12:30
Adam Wyner, Wim Peters, *Semantic Annotations for Legal Text Processing using GATE Teamware*

12:30-13:00 – Position Papers Session

12:30–12:45
Paulo Quaresma, *Legal Information Extraction ← Machine Learning Algorithms + Linguistic Information*

12:45–13:00
Adam Wyner, *Problems and Prospects in the Automatic Semantic Analysis of Legal Texts*

13:00 – 14:00 Lunch break

14:00-15:45 – Session on the Results of the "First Shared Task on Dependency Parsing of Legal Texts"

14:00 –14:20

Felice Dell'Orletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, Giulia Venturi, *The SPLeT--2012 Shared Task on Dependency Parsing of Legal Texts*

14:20 –14:50
Giuseppe Attardi, Daniele Sartiano and Maria Simi, *Active Learning for Domain Adaptation of Dependency Parsing on Legal Texts*

14:50 –15:10
Alessandro Mazzei, Cristina Bosco, *Simple Parser Combination*

15:10–15:30
Niklas Nisbeth, Anders Søgaard, *Parser combination under sample bias*

15:30 –15:45 – Discussion of Shared Task Results

15:45 –16:00 – Closing Remarks

Workshop Organizers

Enrico Francesconi	Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR, Florence, Italy
Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Wim Peters	Natural Language Processing Research Group, University of Sheffield, UK
Adam Wyner	Department of Computer Science, University of Liverpool, UK

Workshop Programme Committee

Kevin Ashley	University of Pittsburgh, USA
Johan Bos	University of Groningen, The Netherlands
Danièle Bourcier	Humboldt Universität, Berlin, Germany
Jack Conrad	Thomson-Reuters, USA
Matthias Grabmair	University of Pittsburgh, USA
Antonio Lazari	Scuola Superiore S.Anna, Pisa, Italy
Alessandro Lenci	Dipartimento di Linguistica, Università di Pisa, Italy
Leonardo Lesmo	Dipartimento di Informatica, Università di Torino, Italy
Thorne McCarty	Reutgers University, USA
Raquel Mochales Palau	Nuance International, Belgium
Paulo Quaresma	Universidade de Évora, Portugal
Tony Russell-Rose	UXLabs, UK
Erich Schweighofer	Universität Wien, Rechtswissenschaftliche Fakultät, Wien, Austria
Rolf Schwitter	Macquarie University, Australia
Manfred Stede	University of Potsdam, Germany
Daniela Tiscornia	Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR, Florence, Italy
Tom van Engers	Leibniz Center for Law, University of Amsterdam, The Netherlands
Giulia Venturi	Scuola Superiore S.Anna, Pisa, Italy
Vern R. Walker	Hofstra University School of Law, Hofstra University, USA
Stephan Walter	Germany
Radboud Winkels	Leibniz Center for Law, University of Amsterdam, The Netherlands

SPLeT-2012 Shared Task Organizers

Felice Dell'Orletta	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Simone Marchi	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Barbara Plank	DISI, University of Trento, Italy
Giulia Venturi	Scuola Superiore S.Anna, Pisa, Italy

Table of contents

Author Index

Preface

The legal domain represents a primary candidate for web-based information distribution, exchange and management, as testified by the numerous e-government, e-justice and e-democracy initiatives worldwide. The last few years have seen a growing body of research and practice in the field of Artificial Intelligence and Law which addresses a range of topics: automated legal reasoning and argumentation, semantic and cross-language legal information retrieval, document classification, legal drafting, legal knowledge discovery and extraction, as well as the construction of legal ontologies and their application to the law domain. In this context, it is of paramount importance to use Natural Language Processing techniques and tools that automate and facilitate the process of knowledge extraction from legal texts.

Since 2008, the SPLeT workshops have been a venue where researchers from the *Computational Linguistics* and *Artificial Intelligence and Law* communities meet, exchange information, compare perspectives, and share experiences and concerns on the topic of legal knowledge extraction and management, with particular emphasis on the semantic processing of legal texts. Within the Artificial Intelligence and Law community, there have also been a number of dedicated workshops and tutorials specifically focussing on different aspects of semantic processing of legal texts at conferences such as JURIX-2008, ICAIL-2009, ICAIL-2011, as well as in the International Summer School "Managing Legal Resources in the Semantic Web" (2007, 2008, 2009, 2010, 2011).

To continue this momentum and to advance research, a 4th Workshop on "Semantic Processing of Legal Texts" was organized at the LREC-2012 conference to bring to the attention of the broader Language Resources/Human Language Technology community the specific technical challenges posed by the semantic processing of legal texts and also share with the community the motivations and objectives which make it of interest to researchers in legal informatics. The outcome of these interactions advance research and applications and foster interdisciplinary collaboration within the legal domain.

New to this edition of the workshop were two sub-events which were meant to provide common and consistent task definitions, datasets, and evaluation for legal-IE systems along with a forum for the presentation of varying but focused efforts on their development.

The first sub-event was a shared task specifically focusing on dependency parsing of legal texts: although this is not a domain-specific task, it is a task which creates the prerequisites for advanced IE applications operating on legal texts, which can benefit from reliable pre-processing tools. For this year our aim was to create the prerequisites for more advanced domain-specific tasks (e.g. event extraction) to be hopefully organized in future SPLeT editions. The languages dealt with have been Italian and English.

The second sub-event was an online, manual, collaborative, semantic annotation exercise, the results of which are presented and discussed at the workshop. The goals of the exercise were: (1) to gain insight on and work towards the creation of a gold standard corpus of legal documents in a cohesive domain; and (2) to test the feasibility of the exercise and to get feedback on its annotation structure and workflow. For this exercise, the language was English.

The workshop and sub-events provided an overview of the state-of-the-art in legal knowledge extraction and management, presented new research and development directions and emerging

trends, and in general furthered the exchange of information regarding legal language resources and human language technologies and their applications.

The papers from the workshop and sub-events are contained in these proceedings.

We would like to thank all the authors for submitting their research and the members of the Program Committee for their careful reviews and useful suggestions to the authors. We also would like to thank the LREC 2012 Organising Committee that made this workshop possible.

The Workshop Chairs

Enrico Francesconi
Simonetta Montemagni
Wim Peters
Adam Wyner

Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts

Giulia Venturi

Scuola Superiore Sant'Anna di Studi Universitari e di Perfezionamento - Pisa (Italy)
Piazza Martiri della Libertà, 33
giulia.venturi@ilc.cnr.it

Abstract

Methodological issues concerning the design and the development of TEMIS, a syntactically and semantically annotated corpus of Italian legislative texts, are presented and discussed in the paper. TEMIS is a heterogeneous collection of texts exemplifying different sub–varieties of Italian legal language, i.e. European, national and local texts. The whole corpus has been dependency annotated and a subset has been enriched with frame–based information by customizing the formalism of the FrameNet project. In both cases, a number of domain–specific extensions of the annotation criteria developed for the general language has been foreseen. The interest in building such a corpus stems from the increasing need for annotated collections of domain–specific texts recognized by both the Artificial Intelligence and Law (AI&Law) community and the Natural Language Processing (NLP) one. In two research communities the benefits of having a resource where both domain–specific content and its underlying linguistic structure are made explicit and aligned are widely acknowledged. To the author knowledge, this is the first annotated corpus of legal texts overtly devoted to be used for legal text processing applications based on NLP tools.

Keywords: Legal Text Processing, Syntactic and Semantic Annotation, Domain–specific Gold Corpora

1. Introduction

This paper presents issues and challenges encountered in designing and developing a corpus of Italian legislative texts enriched with two different layers of linguistic annotation, i.e. a syntactic dependency layer and a semantic one. The interest in building such a corpus stems from the increasing need for annotated collections of domain–specific texts recognized by both the Artificial Intelligence and Law (AI&Law) community and the Natural Language Processing (NLP) one.

On the one hand, in the last few years, a growing body of research and practice has been concerned with the use of Human Language Technology (HLT) for automating knowledge extraction from legal texts and for processing legal language. Such an interest is testified by several events organised on this topic, e.g. the Workshop "Applying Human Language Technology to the Law" held in 2011 (AHLTL 2011)[1] or past editions of the Workshop "Semantic Processing of Legal Texts" (SPLeT), focussed on issues and challenges concerning the use of Natural Language Processing tools in the legal domain.

However, researchers addressing these topics have to confront with the lack of large annotated legal text corpora to be used as reference domain–specific resources. As demonstrated by the promising results achieved in the bio–medical field, corpora annotated at different levels of analysis (e.g. syntactic and semantic levels) play a key role for a number of domain–specific NLP tasks (e.g. biological text mining, the construction of domain–specific ontological resources, event extraction) grounded on the automatic processing of domain–specific corpora.

On the other hand, in the NLP community, it is well known that annotated corpora are valuable resources for the automatic construction of statistical models which can be used in a number of different NLP tasks. However, currently available statistically trained NLP tools are mostly based on corpora made up of texts from the news domain. Applying these tools to out–of–domain corpora is known to be problematic (Gildea, 2001): when applied to domain–specific texts (e.g. bio–medical literature, law texts) their accuracy decreases significantly. Since available domain–specific resources are fundamental in supervised scenarios to adapt statistical NLP tools to new domains, some effort has been devoted to the construction of such resources. The most notable case is represented by the bio–medical domain where the GENIA corpus (Ohta et al., 2002), a collection of biomedical literature annotated with various levels of linguistic (e.g. morphological, syntactic) and semantic information (e.g. domain–specific entities, relational information), has been developed. The corpus is currently used for domain–specific semantic processing applications, e.g. for mining biomedical events from literature (Kim et al., 2008), as well as for supervised domain adaptation purposes, e.g. for improving the performance of statistical syntactic parsers by using bio–medical texts as additional training data (McClosky and Charniak, 2008).

Given these premises, the present article aims at illustrating the main methodological issues faced in syntactically and semantically annotating the TEMIS corpus (Syn**T**actically and S**EM**antically Annotated **I**talian Legislative Corpu**S**). To the author knowledge, it is the first multi–level annotated corpus of legal texts specifically designed to contribute to the development of NLP–based legal text processing applications.

The paper is organized as follows. In Section 2., motivations for developing a multi–level annotated corpus of legal texts for semantic processing purposes are reported, together with related studies. In Section 3., a description

[1] http://wyner.info/research/Papers/AHLTL2011Papers.pdf

of the TEMIS corpus is presented, including some main legal language peculiarities characterizing the corpus with respect to a corpus of general Italian language. Sections 4. and 5. are respectively devoted to illustrate the syntactic and semantic approaches adopted for corpus annotation. Some of the ongoing research activities devoted to use TEMIS for parsing legal texts are illustrated in Section 6. Conclusions and future developments of this work are reported in Section 7.

2. Motivations and related work

The interest in developing a corpus enriched with syntactic and semantic information stems from the acknowledged benefits of having a domain–specific document collection where both domain–specific content and its underlying linguistic structure are made explicit and aligned. In other words, the present work was motivated by the fundamental role that a syntactically and semantically annotated corpus of legal texts could play for several NLP–based applications in the legal domain.
In what follows, recent related studies focussed on developing syntactically or semantically annotated corpora of legal texts are discussed.

2.1. Related work on syntactic annotation of legal texts

This work starts from the idea that any legal text semantic processing application "would be further supported with the creation of a large scale corpus of parsed legal documents" (Wyner and Peters, 2011). Similarly to open–domain NLP–based applications such as Information Extraction, Question Answering, Machine Translation, etc., it is broadly acknowledged that several domain–specific semantic processing applications, e.g. rule extraction from regulations (Wyner and Peters, 2011), formal representations of individual sentences occurring in legal provisions (de Maat and Winkels, 2011), automatic detection of arguments in legal texts (Palau and Moens, 2011), can benefit significantly from operating against the output of a syntactic parser.
However, to the author knowledge, no syntactically annotated corpora of legislative texts are available so far for any language. Accordingly, current statistical parsers are trained on corpora of newspapers, representative of *open–domain* texts. This affects their performances with respect to legal texts.
One exception is the portion of the Turin University Treebank (TUT)[2], developed at the University of Torino, including a section of the Italian Civil Law Code (28,048 word tokens, for a total of 1,100 sentences) annotated with syntactic dependency information. However, this corpus is representative of a legal language sub–variety acknowledged to be less complex with respect to other kinds of legislative texts such as laws, decrees, regulations, etc. According to one of the main scholar of legal language such as Garavelli (2001), the Civil Law Code articles are less representative of the much cited linguistic complexity of Italian *legalese* with respect to other kinds of legislative texts.

[2]http://www.di.unito.it/∼tutreeb/

From an applicative point of view, this is witnessed by the results achieved in the "Dependency Parsing" track of Evalita 2011 (Bosco and Mazzei, 2012) where all participant parsers have shown better performances when tested on the Italian Civil Law Code test set than when tested on newspapers test corpus. On the contrary, the "Domain Adaptation Track" organized at Evalita 2011 (Dell'Orletta et al., 2012), where a sub–set of the corpus presented in this paper has been used, revealed that parsing systems need to be further adaptated to reliably analyse legal texts such as laws, decrees, regulations, etc.

2.2. Related work on semantic annotation of legal texts

Attention to issues and challenges posed by the semantic annotation of legal texts originates from the increasing interest in legal knowledge management tasks based on automatic text processing. Accordingly, several NLP–oriented works have appeared on this topic. Even though they differ in the approach, they aim at making legal texts structured and informative for different automatic semantic processing applications, such as legal argumentation mining (Palau and Moens, 2011), legal text summarization (Hachey and Grover, 2006), court decisions structuring (Kuhn, 2010), legal metadata extraction (see among others for the Italian case (Bartolini et al., 2004; Mazzei et al., 2009; Spinosa et al., 2009)), legal definitions extraction (Walter, 2009), legal case elements and case factors extraction (Wyner, 2010; Wyner and Peters, 2010b; Wyner and Peters, 2010a), legal information retrieval (Maxwell et al., 2009), rule extraction from regulations (Wyner and Peters, 2011), etc.

However, in spite of this widespread interest very little work has been devoted so far to developing a semantically annotated corpus to be used as reference corpus for some of the above mentioned legal text processing applications. To the author knowledge, two exceptions are the Vaccine/Injury Project Corpus (Walker et al., 2011) and the corpus of Brazilian court decisions and legislative texts semantically annotated according to Frame Semantics principles (Bertoldi and Chishman, 2012).

In the first case, Walker and colleagues have built a collection of legal decisions awarding or denying compensation for health injuries allegedly due to vaccinations and they have annotated it with models of the logical structure of the reasoning of the factfinders. The corpus is meant to provide "useful data for formal and informal logic theory, for natural–language research in linguistics, and for artificial intelligence research in those cases". In the second case, a corpus representative of the Brazilian legal language has been annotated with semantic frames information, i.e. by applying the *Frame Semantics* theory (Fillmore, 1985) and the FrameNet paradigm to the semantic annotation of legal texts. The Bertoldi and Chishman' initiative "is part of a larger project that researches how linguistic information could be used to improve legal information management and legal information retrieval in the Brazilian courts".

3. TEMIS: a Syntactically and Semantically Annotated Italian Legislative Corpus

This section is intended to provide the overall description of the TEMIS resource and the principles which guided its design and construction.

The TEMIS corpus has been originally developed in the framework of the author's Ph.D thesis. Starting from a small set of sentences exemplifying legal language, the corpus has been further enlarged in the occasion of the Evalita 2011 campaign where a subset has been used in the "Domain Adaptation" track. As discussed in (Dell'Orletta et al., 2012) where the results of the track were reported, in that occasion the TEMIS subset was used as test corpus. This allowed quantifying the negative impact that the language used in legislative texts such as laws, decrees, regulations, etc. has on the performances of participant parsers trained or developed on newspaper language.

Three annotators, all with graduate training in linguistics, participated both in the syntactic and in the semantic annotation stage.

3.1. Corpus composition

TEMIS is a collection of legislative texts enacted by three different releasing agencies, i.e. European Commission, Italian State and Piedmont Region, and regulating a variety of domains, ranging from environment, human rights, disability rights to freedom of expression. It is a heterogeneous document collection including legal acts such as national and regional laws, European directives, legislative decrees, etc., as well as administrative acts, such as ministerial circulars, decision, etc.

This heterogeneous nature makes TEMIS a resource able to exemplify different sub–varieties of Italian legal language. Table 1 reports how the three different legal text types (i.e. European, national and local texts) are variously represented in the corpus.

Releasing agency	No. tokens	No. sentences
European Commission	6,683	275
Italian State	3,670	94
Piedmont Region	5,453	135
Total	15,804	504

Table 1: Distribution of different legal text types in TEMIS.

3.2. Corpus linguistic profile

In order to get evidence of the linguistic specificity of the legislative texts included in TEMIS, the corpus has been investigated with respect to a number of different parameters, which according to the literature on register variation (Biber and Conrad, 2009) are indicative of textual genre differences.

Different kinds of features have been taken here as representative of the linguistic profile of the considered legislative texts. They range from raw text features, such as sentence length, to more complex ones (e.g. parse tree depth) detected from the syntactic level of annotation. In what follows the most significant ones are illustrated and discussed.

A comparison with the respective features for an Italian newswire corpus, chosen to be representative of general Italian language, helps to highlight the TEMIS's main linguistic characteristics. The ISST–TANL corpus, jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project, has been used. The corpus consists of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.).

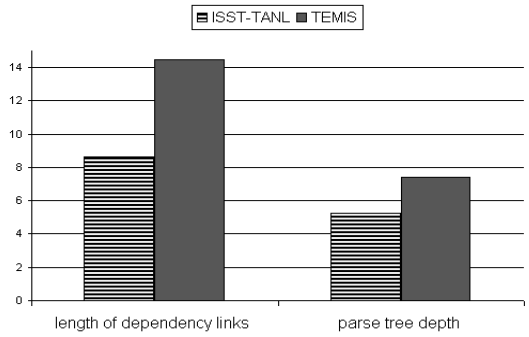
The TEMIS and ISST–TANL corpora differ significantly in many aspects starting from the average sentence length, calculated as the average number of words per sentence. As Table 2 shows, some differences can be also found amongst the three considered kinds of legal text sub–varieties. Notably, the legal texts enacted by the European Commission (TEMIS–EU, in the Table) show a behaviour which is more similar to ordinary language than the national (TEMIS–NAT) and local (TEMIS–LOC) legal texts.

Corpus	Avg sentence length (in tokens)
ISST–TANL	21.87
TEMIS	31.36
TEMIS–EU	24.56
TEMIS–NAT	39.04
TEMIS–LOC	41.95

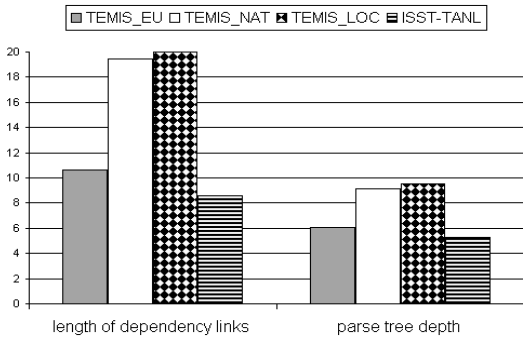
Table 2: Average sentence length in *i)* TEMIS and ISST–TANL corpora and *ii)* the three TEMIS's sub–corpora.

Interestingly, several differences can be found between the two corpora with respect to the distribution of features typically correlated with text complexity, such as parse tree depth and length of dependency links. According to the approach to linguistic monitoring described in (Dell'Orletta et al., 2011), the TEMIS and ISST–TANL corpora have been compared with respect to *i)* the average length of dependency links, measured in terms of the words occurring between the syntactic head and the dependent, and *ii)* the average depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf. As it can be seen in Figure 1(a), *i)* legislative sentences contain dependency links much longer on average (14.5) than the ones of the general–Italian sentences (8.61) and *ii)* the average parse tree height of TEMIS (7.44) is higher than the one characterizing the ISST–TANL sentences (5.28). In addition, as it was previously pointed out, the Italian European legal texts have syntactic features which make them more similar to ordinary language than the national and local legal texts (see Figure 1(b)).

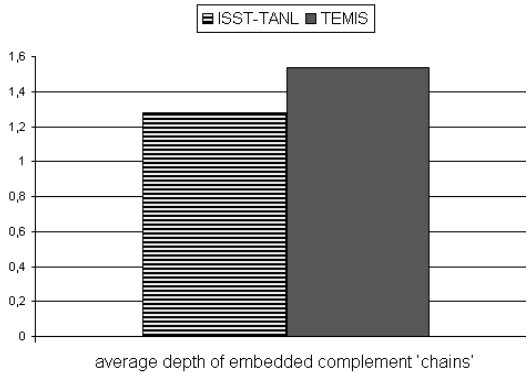
It is here worth noting (see Figure 1(c)) that TEMIS's sentences are characterized by an average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers (1.54) higher than the one of ISST–TANL's sentences (1.28). However, the crucial distinguishing characteristic of legislative sentences appears to be the different percentage distributions of embedded complement



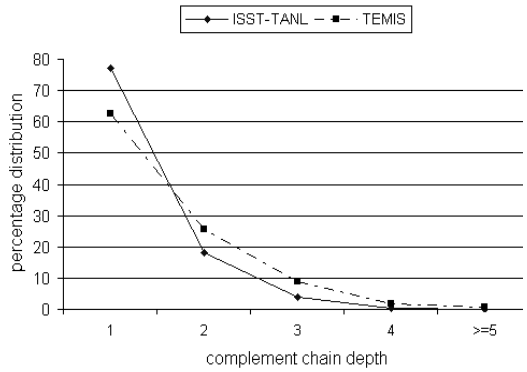
(a) Average length of dependency links and average depth of parse tree in TEMIS and ISST–TANL.



(b) Average length of dependency links and average depth of parse tree in TEMIS–EU, –NAT, –LOC and ISST–TANL.



(c) Average depth of embedded complement 'chains' in TEMIS and ISST–TANL.



(d) Percentage distributions of embedded complement 'chains' by depth in TEMIS and ISST–TANL.

Figure 1: Comparative syntactic behaviours in TEMIS and ISST–TANL.

'chains' by depth. As Figure 1(d) shows, legislative texts appear to have an higher percentage of deep complement 'chains' with respect to the Italian reference corpus.

4. The syntactic level of annotation

All the 504 sentences included in the TEMIS corpus have been enriched with a syntactic level of annotation. For this purpose, a semi–automatic strategy has been adopted. Firstly, the TEMIS corpus was automatically dependency-parsed by the DeSR parser (Attardi, 2006) using *i)* Support Vector Machine as learning algorithm and *ii)* the ISST–TANL corpus as training corpus.

Secondly, the result of the first stage was manually revised. The "Dependency Grammar Annotator" (DgAnnotator) tool[3] was used for such a manual revision step.

This semi–automatic annotation strategy is meant to reduce the annotation arbitrariness to a minimum. This allowed to keep consistent the manual annotation as well as to identify the main parsing errors due to the unique features of legal language.

[3]It is an annotating and visualizing Dependency Graphs tool freely available at http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

4.1. Dependency Annotation Scheme and Data Format

The dependency syntactic annotation scheme developed for the ISST–TANL corpus labelling has been used for the TEMIS corpus annotation. That is, the dependency tagset[4] was maintained even though a number of extensions of the annotation criteria have been introduced in order to properly handle legal language syntactic peculiarities.

The dependency annotation format adheres to the standard CoNLL–2007 tabular format used in the "Shared Task on Dependency Parsing" (Nivre et al., 2007). Accordingly, each word–token is provided with information concerning the corresponding lemma, coarse- and fine–grained part–of–speech[5], morphological features[6], head of the dependency relation and the dependency relation type.

[4]A description of the dependency tagset can be found at http://poesix1.ilc.cnr.it/ISST-TANL-DEPtagset-web.pdf

[5]A description of the part-of-speech (coarse- and fine-grained) tagsets can be found at http://poesix1.ilc.cnr.it/ISST-TANL-MStagset-web.pdf

[6]A description of the part-of-speech (coarse- and fine-grained) tagsets can be found at http://poesix1.ilc.cnr.it/ISST-TANL-MS_FEATStagset-web.pdf

For example, the following sentence is annotated as Table 3 reports:

- *Gli Stati membri provvedono affinché il gestore sia obbligato a trasmettere all'autorità competente una notifica entro i seguenti termini.* ('Member States shall require the operator to send the competent authority a notification within the following time–limits'.)

In Table 3, it can be noted that each word form (in the column headed *FORM*) univocally marked by a numerical identifier (column *ID*) is associated with its corresponding lemma (column *LEMMA*), its coarse– (column *CPOSTAG*) and fine–grained (column *POSTAG*) part–of–speech and its morphologycal treats (column *FEATS*). Moreover, the annotation makes explicit the head of the dependency syntactic relation in which each word is involved (column *HEAD*) and the type of dependency relation (columnn *DEPREL*). For example, Table 3 shows that the word *notifica* ('notification') is the object (*obj*) of the verb *trasmettere* ('send').

4.2. Domain–specific extensions of the open–domain annotation criteria

The annotation criteria developed for the annotation of an *open–domain* corpus such as the ISST–TANL needed to be extended in order to properly handle specific syntactic peculiarieties specific to the legal language. Such extensions are concerned with different levels of text annotation ranging from the sentence splitting to the dependency annotation. The most significant cases are described in the following sections.

4.2.1. Sentence splitting

Differently from the criteria adopted for the open–domain case, here sentence splitting was overtly meant to preserve the original structure of the legal text. This entails that also punctuation marks such as ';' and ':', when followed by a carriage return, are treated as sentence boundary markers. Such an extension allowed to handle with specific cases frequently occurring in legislative texts, such as:

1. sentences that, occurring in a legislative preamble, start with phrases, such as *considerato che* ('Having regard to'), and end with a clause boundary punctuation mark, such as ';'

2. sentences that end with a clause boundary punctuation mark such as ':' and introduce an itemized list

3. sentences that, part of an itemized list, end with a clause boundary punctuation mark such as ';'.

4.2.2. Dependency annotation

In order to successfully cope with domain–specific syntactic constructions hardly or even never occurring in the ISST–TANL corpus, dependency annotation criteria have been extended to cover the annotation of the main following cases:

1. elliptical constructions frequently adopted in citations to whole legal texts or to specific partitions of legal texts (e.g. article, paragraph, etc.). This is the case, for example, of the sentence *i decreti legislativi di cui al comma 1 si conformano ...* ('legislative decrees referred to in paragraph 1 shall comply ...') which has been annotated as Figure 2 shows. Since a verb is missing in the relative clause *di cui al comma 1* ('referred to in paragraph 1'), a relative–modifier dependency relation (*mod_rel*) has been found between the antecedent *decreti* ('decrees'), the head token, and *al* ('to'), the dependent token;

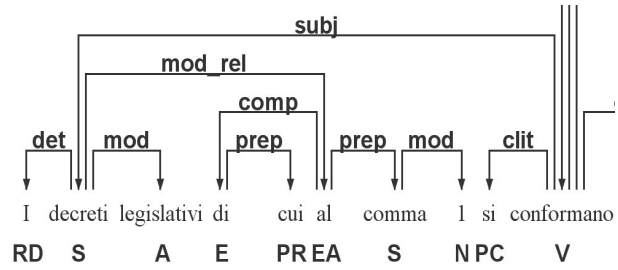


Figure 2: Example of annotation of elliptical construction.

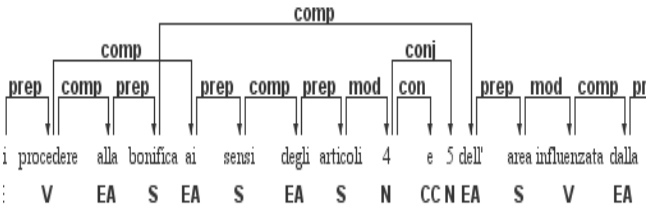


Figure 3: Example of non–projective link.

2. participial phrases, such as *fatto salvo* ('without any reserve'), used to express exceptions or limitations to main clauses. For example, as Figure 5 shows, in the sentence *il contravventore, fatti salvi ogni altro adempimento o comminatoria previsti dalle leggi vigenti, è tenuto al pagamento di una sanzione amministrativa ...* ('the infringer, without prejudice to any other obligation or on pain of applicable law, is required to pay an administrative penalty'), a modifier dependency relation (*mod*) has been found between the head of the participial phrase (i.e. *fatti*) and the syntactic head of the main clause (i.e. *tenuto* 'required');

3. non–projective links, often occurring in legal texts and mostly due a frequent exploitation of the free–word order nature of the Italian language. This is the case of the annotation excerpt of the sentence *E' fatto comunque salvo l'obbligo di procedere alla bonifica ai sensi degli articoli 4 e 5 dell'area influenzata dalla fonte inquinante* ('Is without prejudice to the obligation to carry out drenage in accordance with articles 4 and 5 of the area affected by pollution sources') reported in Figure 3 where not all the tokens part of the *comp* relation are dependent from the head token *bonifica* ('drenage'); on the contrary, the head of the token *ai* ('in') is *procedere* ('carry out');

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Gli	il	R	RD	num=p|gen=m	2	det
2	Stati	Stati	S	SP	_	4	subj
3	membri	membro	S	S	num=p|gen=m	2	mod
4	provvedono	provvedere	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
5	affinché	affinché	C	CS	_	4	mod
6	il	il	R	RD	num=s|gen=m	7	det
7	gestore	gestore	S	S	num=s|gen=m	9	subj_pass
8	sia	essere	V	VA	num=s|per=3|mod=c|ten=p	9	aux
9	obbligato	obbligare	V	V	num=s|mod=p|gen=m	5	sub
10	a	a	E	E	_	9	arg
11	trasmettere	trasmettere	V	V	mod=f	10	prep
12	all'	a	E	EA	num=s|gen=n	11	comp_ind
13	autorità	autorità	S	S	num=n|gen=f	12	prep
14	competente	competente	A	A	num=s|gen=n	13	mod
15	una	una	R	RI	num=s|gen=f	16	det
16	notifica	notifica	S	S	num=s|gen=f	11	obj
17	entro	entro	E	E	_	11	comp_temp
18	i	il	R	RD	num=p|gen=m	20	det
19	seguenti	seguente	A	A	num=p|gen=n	20	mod
20	termini	termine	S	S	num=p|gen=m	17	prep
21	.	.	F	FS	_	4	punc

Table 3: An example of an annotated sentence in CoNLL format extracted from TEMIS.

4. internal partitions of a legislative text (e.g. article, paragraph) that are hierarchically organized. They are treated as embedded modifier 'chains' governed by a nominal head, as exemplified by the annotation of the sentence *ai sensi dell'articolo 94, comma 3, lettera a) della l.r. 44/2000* ('under article 94, paragraph 3, letter a) of the local act 44/2000') reported in Figure 4. In this case, the internal partitions of the *l.r. 44/2000* ('local act 44/2000'), i.e. *articolo 94, comma 3, lettera a)* ('article 94, paragraph 3, letter a)'), has been annotated as a chain of nominal modifiers (*mod*).

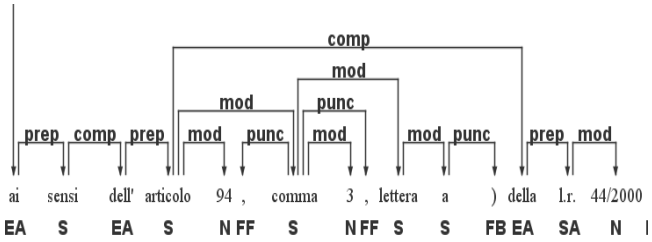


Figure 4: Example of annotation of internal partitions of a legislative text.

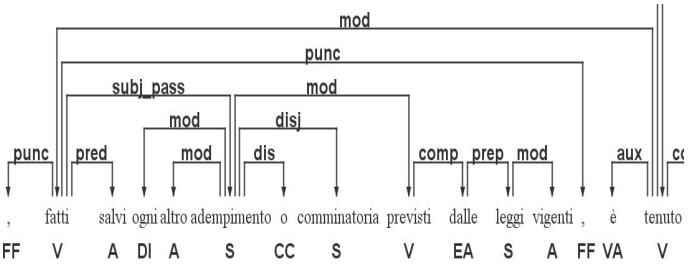


Figure 5: Example of annotation of participial phrase.

5. The semantic level of annotation: a FrameNet–based approach

A subset of the TEMIS corpus has been further enriched with a level of semantic annotation. The semantic annotation paradigm developed in the framework of the FrameNet project[7] has been adopted and specialized in order to properly describe the lexical semantic content of legislative texts included in TEMIS.

This initiative is part of the work that has been jointly done at the University of Pisa (Department of Linguistics) and at the Institute of Computational Linguistics "Antonio Zampolli" (ILC-CNR) where a frame-based annotation of the ISST-TANL corpus has been carried out in order to enrich the treebank with semantic frame information, as described in (Lenci et al., 2012).

By starting from the suggestion expounded by Dolbey et al. (2006a) that FrameNet can be seen "as a backbone of several domain–specific FrameNets", the annotation methodology adopted for the ISST–TANL corpus has been extendend and specialized. This aims at showing that a FrameNet–like approach to text annotation can be suitable for insightful analyses of general language, as well as an advantageous starting point for producing descriptions of syntactic and semantic combinatorial possibilities exhibited in a specialized language such as the legal language.

In particular, the annotation effort has been devoted to making explicit how the three main deontic modalities, i.e. *obligation, permission, prohibition*, are linguistic realized

[7] https://framenet.icsi.berkeley.edu/fndrupal/

in the TEMIS corpus. Accordingly, the annotation of the frames reported in Table 4 has been the main focus.
Formally represented in the existing legal ontologies, these three fundamental legal concepts are hardly associated with their actual lexical realization. Accordingly, the underlying idea here is that a FrameNet–based and linguistically–oriented representation of legal semantics would complement a domain–oriented one by providing a semantic description anchored to its corresponding textual realization. As Dolbey et al. (2006a) suggested, "FrameNet–style ontological descriptions of language can be integrated with information from" already existing domain–specific ontologies, such as bio–medical ontologies (in Dolbey's case) or legal ontology.

5.1. The FrameNet project

FrameNet is a lexical resource for English, based on *Frame Semantics* (Fillmore, 1985) and supported by corpus–evidence. The goal of the FrameNet project is to document the range of semantic and syntactic combinatory possibilities of each word in each of its senses. Typically, each sense of a word belongs to different Semantic Frame, conceived in (Ruppenhofer et al., 2010) as "a script–like conceptual structure that describes a particular type of situation, object or event along with its participants and properties". For example, the APPLY_HEAT frame describes a common situation involving participants such as "Cook" and "Food", etc. , called Frame Elements (FEs), and is evoked by Lexical Units (LUs) such *bake, blanch, boil, broil, brown, simmer*, etc. As shown by the following example, the frame–evoking LU can be a verb (bolded in the example) and its syntactic dependents (those written in subscript) are its FEs:

- [Matilde $_{Cook}$] **fried** [the catfish $_{Food}$] [in a heavy iron skillet $_{Heathing\_instrument}$].

The type of representation produced by FrameNet is a network of "situation–types" (frames) organized across inheritance relations between frames, as opposed to a network of meaning nodes, as in the case of WordNet. In FrameNet, FE can be also specified with Semantic Types (i.e. ontological categories) employed to indicate the basic typing of fillers that are expected in the FE. Most of these semantic types correspond directly to synset nodes of WordNet, and can be mapped onto already existing ontologies. FrameNet currently contains more than 1,123 frames, covering 12,280 Lexical Units; these are supported by more than 188,778 FrameNet–annotated example sentences.
Despite the fact that FrameNet annotations are triples where each FE realization is coupled with its phrase type (e.g. NP, PP, etc.) and its grammatical function (e.g. object, subject, etc.) played in the annotated sentence, as overtly claimed by (Dolbey, 2009), "FrameNet annotations are not linked to syntactic parse trees"; consequently, it is often the case that frame elements instantiations do not "correspond to syntactic constituents provided by a syntactic parse of the whole sentence".

5.2. Annotation methodology

The TEMIS's semantic annotation has been focussed on a sub–set of the whole corpus. Namely, a set of 9,273 tokens (for a total of 226 sentences) extracted from the three considered sub–varieties of legislative texts has been semantically annotated.
The annotation methodology which has been followed intends to continue the one described in (Venturi, 2011b) where a FrameNet–based approach to the semantic annotation of Italian legislative texts has been presented. The strategy devised for the annotation of the ISST-TANL corpus has been mostly adopted, even though a number of domain–specific customizations have been considered.
Similarly to the ISST–TANL case, the annotation was carried out manually with the SALTO tool (Erk et al., 2003). The syntactic annotation level of the TEMIS corpus was first automatically converted into the TIGER/SALSA XML format, and then loaded onto SALTO, together with the ontology of semantics frames and FEs derived from FrameNet, by following the methodology fully described in (Lenci et al., 2012).
The annotation has been carried out on top of dependency syntactic representation. This entails that during the annotation phase frames and FEs have been anchored to the dependency annotation. Such an approach mostly resembles the strategy to frame–semantic corpus annotation that the SALSA project (Burchardt et al., 2006) adopted, which included manually annotating a large corpus of German newspapers with semantic role information starting from a syntactically annotated corpus.
FrameNet–style annotations are not linked to syntactic parse trees: the choice to ground the level of semantic annotation on the fully parsed texts is intended to overcome this state of affairs. Indeed, as Dolbey (2009) emphasizes, the annotation strategy adopted in the FrameNet project may cause "difficulties for end users who want to perform automatic processing that includes information from FrameNet's annotation collection".
An example of annotation is reported in Figure 6, where the following two frames have been annotated on top of the syntactic dependency tree of the sentence *Obbligati al pagamento della tassa sono gli esercenti i grandi impianti di combustione di cui all'articolo 1* (lit. 'Obligated to tax payment are tradespeople of big combustion plants mentioned by Article 1'):

1. BEING_OBLIGATED frame, evoked by the verb *obbligare* ('to obligate') in the passive form (i.e. *essere obbligato*, 'be obligated'), with the FEs "Responsible party" and "Duty", instantiated as passive subject (*subj_pass*) and as complement (*comp*) of the verb, respectively;

2. COMMERCE_PAY frame, evoked by the deverbal noun *pagamento* ('payment'), with the constructionally null instantiated FE "Buyer", omitted as the passive subject of the main verb *obbligare* ('to obligate'), and the "Money" FE, syntactically instantiated as a complement (*comp*) of the deverbal noun.

In order to properly represent legal semantics, the foreseen customizations of the annotation methodology adopted for the ISST–TANL corpus mainly concern the *i)* the kind of legal content one would like to make explicit and *ii)* the

Deontic modality: **obligation**		
FrameNet frame	FrameNet definition	Frame–evoking LUs annotated in TEMIS
OBLIGATION_SCENARIO (Non–Lexical Frame)	Under some, usually implicit, Condition a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable social Consequence for the Responsible_party. This Consequence may or may not be stated overtly.	–
BEING_OBLIGATED	Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly.	*tenuto* 'required', *obbligato* 'obligated', *chiamato* 'called', *obbligo* 'obligation', *costretto* 'forced', *sottoposto* 'subjected', *(essere)soggetto* '(to be)subject', *(avere)obbligo* '(have)obligation'
BEING_OBLIGATORY	Under some Condition, usually left implicit, a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable Consequence for the Responsible_party, which may or may not be stated overtly. Compare this frame to the Being_obligated frame.	*obbligatorio* 'obligatory', *spettare* 'to be due', *dovuto* 'due', *incombere* 'to be incumbent'
IMPOSING_OBLIGATION	A Duty is imposed on a Responsible_party according to a Principle which regulates how the Responsible_party should respond to a Situation. The Situation may be expressed metonymically by reference to an Obligator, whose action invokes the Principle. It is only rarely the case that the Principle and the Situation/Obligator are both expressed overtly.	*irrogato* 'imposed', *irrogare* 'to impose', *disporre* 'to decide', *prevedere* 'to provide', *imposto* 'imposed', *predisporre* 'to establish', *definire* 'to fix', *stabilire* 'to establish', *istituire* 'to introduce', *prescrizione* 'prescription', *obbligare* 'to obligate', *disposto* 'provided', *determinare* 'to fix', *(fare)obbligo* '(make)obligation'
Deontic modality: **permission**		
PERMITTING	In this frame a State_of_affairs is permitted by a Principle. Raising constructions are common in this frame. In this frame the Principle which sanctions the State_of_affairs is not an agent who grants permission to a specific individual or group of individuals, and thus differs from the Grantor in the Grant_permission frame.	*autorizzato* 'authorized', *autorizzare* 'to authorize', *ammesso* 'permitted', *accordato* 'granted', *consentire* 'to allow', *consentito* 'allowed', *concessione* 'permission', *concesso* 'granted', permesso 'permission'
Deontic modality: **prohibition**		
PROHIBITING	In this frame a State_of_affairs is prohibited by a Principle. Raising constructions are common in this frame. In this frame the Principle which prohibits the State_of_affairs is not an agent who denies permission to a specific individual or group of individuals, and thus differs from the Authority in the Deny_permission frame.	*interdizione* 'disability', *divieto* 'prohibition', *vietato* 'prohibited', *(fare)divieto* '(make)prohibition'
DENY_PERMISSION	In this frame, an Authority orders a Protagonist not to engage in an Action.	*divieto* 'prohibition', *interdizione* 'disability', *negare* 'to deny', *proibire* 'to prohibit', *(fare)divieto* '(make)prohibition'

Table 4: FrameNet frames describing the *obligation, permission, prohibition* deontic modalities and the corresponding evoking lexical units annotated in the TEMIS corpus.

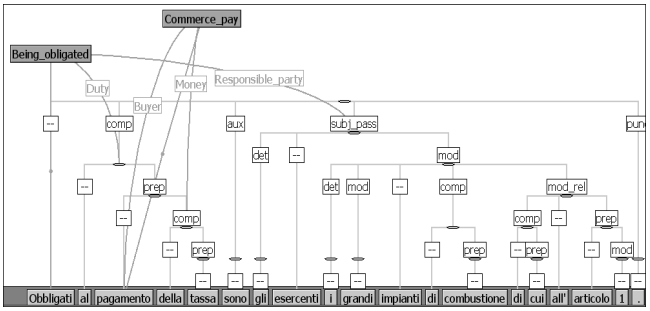


Figure 6: An example of frame–semantic annotation using the SALSA tool.

kind of domain–specific customizations and extensions of the general FrameNet resource required by the legal domain.

5.2.1. Between lexicographic and full–text annotation

The first issue to address for a legal semantic annotation task concerns the kind of text content to make explicit. It is well–known that the law simultaneously *describes* objects and events and *regulates* them. Thus, *legal domain knowledge* is mixed with knowledge of domain of interest to be regulated (i.e. *world knowledge*).

To successfully cope with this mixture, a special annotation mode has been adopted, which is meant to be between the two annotation modes that FrameNet has used, i.e. *lexicographic* and *full–text* annotation. Here, two different annotation strategies have been followed to annotate frame–semantic information evoked by lexical units that convey *legal* knowledge and by those lexical units that express *world* knowledge. To be more specific, the annotation of lexical units that convey fundamental legal concepts, e.g. those units expressing deontic modalities (e.g. *proibire* 'to prohibit', *divieto* 'prohibition', *obbligare* 'to obligate', etc.), followed the lexicographic mode. That is, the annotation started from a list of lexical units belonging to the legal domain. In addition, a full–text annotation mode has been followed in the annotation of frame–semantic information conveyed by those lexical units in the regulated domain. This annotation has been done only when the lexical units were already part of a situation–type (i.e. a semantic frame) belonging to the legal domain. In fact, frame–evoking lexical units that express domain–specific world knowledge have been annotated only when they served as lexical fillers of a FE part of a legal frame.

For example, the sentence given below, where the verb *ha vietato* ('prohibited') evokes a PROHIBITING frame conveying *legal* information, has been annotated as follows:

- [*La decisione 90/200* $_{Principle}$] *ha* [***vietato*** $_{TARGET}$] [*l'esportazione dal Regno Unito di taluni tessuti e organi bovini* $_{State\_of\_affairs}$] [*solo dopo il 9 aprile 1990* $_{Time}$].

  ('[The decision 90/200 $_{Principle}$] [prohibited $_{TARGET}$] [the exportation from the United Kingdom of certain bovine tissues and organs $_{State\_of\_affairs}$] [only after the 9th of April 1990 $_{Time}$]').

The deverbal noun *esportazione* ('exportation') evokes an EXPORTING frame, conveying *world* knowledge, and was included in the FE "State of affairs" which belongs to the PROHIBITING frame. Therefore, a second annotation has been provided for that sentence as follows:

- [*La decisione 90/200* $_{Principle}$] *ha* [*vietato* $_{TARGET}$] [[***l'esportazione*** $_{TARGET}$] [*dal Regno Unito* $_{Exporting\_area}$] [*di taluni tessuti e organi bovini* $_{Goods}$] $_{State\_of\_affairs}$] [*solo dopo il 9 aprile 1990* $_{Time}$].

  ('[The decision 90/200 $_{Principle}$] [prohibited $_{TARGET}$] [[the exportation$_{TARGET}$] [from the United Kingdom$_{Exporting_area}$] [of certain bovine tissues and organs $_{Goods}$] $_{State\_of\_affairs}$] [only after the 9th of April 1990 $_{Time}$]').

5.2.2. Domain–specific customizations issues

The annotation methodology mostly consists of maintaining and reusing the semantic frames and FEs already defined in FrameNet. However, several domain–specific customizations were needed. Three customization strategies,

fully described in (Venturi, 2011a), have been followed. They differ in their increasing degree of modification to the FrameNet resource and they concern:

1. the introduction of one or more FEs within an existing frame. This happened when FrameNet did not foresee that an important piece of information was part of the background knowledge evoked by a predicative lexical unit. For example, FrameNet did not include a "Purpose" FE in the BEING_OBLIGATED frame, even though this piece of information is needed to fully describe the semantics conveyed by this frame, as shown in the following annotated sentence:

   - [*Per la realizzazione delle opere previste nelle convenzioni già assentite alla data del 30 giugno 2002, ovvero rinnovate e prorogate ai sensi della legislazione vigente* $_{Purpose}$] [*i concessionari* $_{Responsible\_party}$] *sono* [***tenuti*** $_{TARGET}$] [*ad appaltare a terzi una percentuale minima del 40 per cento dei lavori,* $_{Duty}$] [*applicando le disposizioni della presente legge ad esclusione degli articoli 7, 14, 19, commi 2 e 2-bis, 27, 32, 33* $_{Condition}$]. (Lit. [For the realization of works planned in the convenctions already assented on the date of the 30th June 2002, that is renewed and extended under the in force law $_{Purpose}$] [the agents $_{Responsible\_party}$] are [**bound** $_{TARGET}$] [to contract out to third party a percentage minimal of the 40% of works, $_{Duty}$] [enforcing the provisions of the present law with the exception of articles 7, 14, 19, paragraphs 2 and 2-bis, 27, 32, 33 $_{Condition}$].)

   This sentence demonstrates that to fully characterize the BEING_OBLIGATED frame for the legal domain it is necessary to account for the particular scope that can be achieved if the "Responsible_party" performs a "Duty" (i.e. the "Purpose");

2. the specification of domain–specific semantic types in order to classify FEs. This is done by adding semantic types taken from an existing legal ontology, when no proper semantic type is available in FrameNet. For example, in the BEING_OBLIGATED frame neither the FE "Duty" nor "Responsible_party" were assigned any semantic type. Therefore, for these FEs the domain–specific customization included the typing with the semantic type 'Duty' and 'Legal Subject' respectively, two classes (i.e. two juridical concepts) which were taken from the Core Legal Ontology (CLO) (Gangemi et al., 2005);

3. the creation of new semantic frame(s). This represents the most controversial kind of customization. As Dolbey et al. (2006b) warns, on the one hand, the introduction of a new frame to specify domain–specific information would result in a richer representation of domain–specific semantics; on the other hand, there would be an increase in the complexity of the network of frames. For example, a new GRANT_LEGAL_PERMISSION frame was added in order to characterize a situation–type where an authority grants a permission to a grantee. In FrameNet there are two different frames that may evoke such a situation: PERMITTING and GRANT_PERMISSION. The first one describes a situation where a "State of Affairs is permitted by a Principle"; the second one represents a situation where "a Grantor (either a person or an institution) grants permission for a Grantee to perform an Action". However, the latter frame, according to FrameNet's definition, "does not include situations where there is a state of permission granted by authority or rule of law". The new suggested frame inherits some of the FEs of the GRANT_PERMISSION frame with a number of domain–specific customizations[8]. Thanks to this newly introduced frame, it is thus possible to properly represent the legal content of the following sentence:

   - [*Il Ministero della sanità* $_{Legal\_grantor}$], *per quanto riguarda gli aspetti ambientali d'intesa con il Ministero dell'ambiente,* [***autorizza*** $_{TARGET}$] [*ai sensi del presente decreto* $_{Circumstances}$] [*l'immissione sul mercato e l'utilizzazione nel territorio italiano di un biocida* $_{Permitted\_action}$]. (Lit. [The Ministry of Health $_{Legal\_grantor}$], regarding the environmental aspects according to the Ministry of Environment [**authorizes** $_{TARGET}$] [under this decree $_{Circumstances}$] [the placing on the market and the usage in Italian territory of a biocidal $_{Permitted\_action}$].)

6. Using TEMIS

As mentioned in Section 3., a subset of TEMIS was used as test corpus in the occasion of the "Domain Adaptation" track at Evalita 2011 to test the performances of statistical parsers trained or developed on newspaper corpora. Similarly, a subset of the syntactically annotated resource described here is currently been used in the "First Shared Task on Dependency Parsing of Legal Texts" at SPLeT 2012.
In this section, the author intends to illustrate the results of first experiments devoted to use TEMIS to parse legal texts. The main aim is to quantify the impact that the language of the legislative texts included in TEMIS has on the accuracy of the parser exploited here.
Results are reported in Table 5. The DeSR parser (Attardi, 2006) has been exploited using *i)* Support Vector Machine as learning algorithm and *ii)* the ISST–TANL corpus as training corpus. The parser has been tested on a test set of 5,165 tokens extracted from the ISST–TANL corpus (ISST–TANL–test) and on a test set of 5,866 tokens extracted from

[8]The foreseen customizations are the following (on the left of the $<$ the FEs of the new GRANT_LEGAL_PERMISSION frame, on the right the corresponding FEs of the GRANT_PERMISSION frame already existing in FrameNet):

- 'Legal grantor' $<$ 'Grantor',
- 'Grantee' $<$ 'Grantee',
- 'Permitted_action' $<$ 'Action'.

the TEMIS corpus (TEMIS–test). Evaluation have been carried out in terms of the standard accuracy dependency parsing measure, i.e. labeled attachment score (LAS): the percentage of tokens for which it has predicted the correct head and dependency relation.

Test set	LAS
ISST–TANL–test	79.71
TEMIS–test	74.72
TEMIS–EU–test	79.30
TEMIS–NAT–test	72.75
TEMIS–LOC–test	72.19

Table 5: Performance of the DeSR parser trained on the ISST–TANL corpus and tested on TEMIS.

As it was expected, the parser trained on the newswire domain has lower performance when tested on the legislative texts of TEMIS. The DeSR performances have a drop of 4.99 percentage points passing from a LAS of 79.71% obtained on the ISST–TANL–test to a LAS of 74.72% on the TEMIS–test.

In order to test the parser accuracy with respect to the three considered legal text sub–varieties, a further experiment has been carried out. Thus, DeSR has been tested on *i)* a test of 1,932 tokens taken from the TEMIS's sub–corpus made up of European Italian legal texts (TEMIS–EU–test), *ii)* a test of 1,971 tokens from the sub–corpus of national legal texts and *iii)* a test of 1,963 tokens from the sub–corpus of local legal texts. Interestingly, the parser has the highest performance when tested on the European legal texts. This is in line with the results of the linguistic monitoring reported in Section 3.2., where it has been demonstrated that this latter legal text sub–variety has a linguistic behaviour which is more similar to newswire texts than the national and local legal texts.

In addition, it is suggested here that the TEMIS corpus can be helpful for domain adaptation purposes. By embracing a supervised approach, it has been used to improve the performance of DeSR as domain–specific additional training data. A pilot experiment has been carried out adding a set of 9,940 tokens from TEMIS to the parser training data, i.e. the training set portion of ISST–TANL. Interestingly, the parser has an improvement of 6.66 percentage points passing from a LAS of 74.72 to a LAS of 81.38.

7. Conclusion and future work

Methodological issues concerning the design and the development of TEMIS, a syntactically and semantically annotated corpus of Italian legislative texts, have been presented and discussed. To the author knowledge, this is the first initiative aiming at building an annotated corpus of legal texts which is overtly devoted to be used for legal text processing applications based on NLP tools. Accordingly, a number of future directions of research can be foreseen.

As illustrated in Section 6., the syntactically annotated resource can be used to parse legal texts as training data of a statistical parser. In addition, it can be exploited in a supervised domain adaptation scenario to improve the performances of a parser originally trained on a different domain.

Currently, it has been planning to increase the amount of sentences semantically annotated in TEMIS by *i)* annotating additional textual instances of the deontic modalities considered so far and *ii)* making explicit further information relevant for the legal domain. The latter direction of research is related to the number of foreseeable legal uses of the corpus. For example, the organization principles of the semantic annotation methodology adopted in the present work could be used to linguistically ground the logical structure of the reasoning of the factfinders in a corpus such as the Vaccine/Injury Project Corpus (Walker et al., 2011). Accordingly, an adequate ontology of frames and frame elements could be devised aiming at associating (semantic) information concerning, for example, under– or over–compensation for health injuries with their (textual) linguistic realization.

Finally, the TEMIS corpus semantically annotated can be a useful resource for several semantic processing tasks, such as Semantic Role Labeling (SRL) of legal texts. Following the strategy adopted for the ISST–TANL corpus which has been recently used as training corpus in the framework of the "Frame Labeling over Italian Texts" (FLaIT) task of Evalita 2011 (Basili et al., 2012), the frame–based information annotated in TEMIS can be exploited to train a domain–specific semantic role labeler.

8. Acknowledgements

The author would like to thank Francesco Asaro and Tommaso Petrolito who have contributed to both the syntactic and the semantic annotation process.

9. References

G. Attardi. 2006. Experiments with a multilanguage non–projective dependency parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X '06)*, pages 166–170, New York City, New York.

R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, and C. Soria, 2004. *Automatic classification and analysis of provisions in legal texts: a case study*, pages 593–604. Lecture Notes in Computer Science, 3292/2004, Springer-Verlag.

R. Basili, A. Lenci, D. De Cao, and G. Venturi. 2012. Frame labeling over italian texts. In *Working Notes of EVALITA 2011*, Rome, Italy.

A. Bertoldi and R. Chishman. 2012. Frame semantics and legal corpora annotation: theoretical and applied challenges. *Linguistic Issues in Language Technology*, 7(9).

D. Biber and S. Conrad. 2009. *Register, genre, and style*. Cambridge, Cambridge University Press.

C. Bosco and A. Mazzei. 2012. The evalita 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*, Rome, Italy.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padò, and M. Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genova, Italy.

E. de Maat and R. Winkels. 2011. Formal models of sentences in dutch law. In *Proceedings of the Workshop Applying Human Language Technology to the Law (AHLT 2011)*, pages 28–40, Pittsburgh, Pennsylvania.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland.

F. Dell'Orletta, S. Marchi, S. Montemagni, G. Venturi, T. Agnoloni, and E. Francesconi. 2012. Domain adaptation for dependency parsing at evalita 2011. In *Working Notes of EVALITA 2011*, Rome, Italy.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006a. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the Biomedical Ontology in Action, Workshop at KR–MED*, pages 87–94, Baltimore, Maryland.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006b. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the "Biomedical Ontology in Action" Workshop at KR-MED*, Baltimore, Maryland.

A. Dolbey. 2009. *BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology*. Ph.D. thesis, University of California, Berkeley.

K. Erk, A. Kowalski, and S. Padò. 2003. The salsa annotation tool-demo description. In *Proceedings of the 6th Lorraine–Saarland Workshop*, pages 111–113, Nancy, France.

C.J. Fillmore. 1985. Frame and the semantics of understanding. *Quaderni di semantica*, IV(2), dicembre:222–254.

A. Gangemi, M.T. Sagri, and D. Tiscornia. 2005. A constructive framework for legal ontologies. In J. Breuker R. Benjamins, P. Casanovas and A. Gangemi, editors, *Law and the Semantic Web*, pages 97–124. Berlin: Springer verlag edition.

B. Mortara Garavelli. 2001. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino, Einaudi.

D. Gildea. 2001. Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 167–202, Pittsburgh, PA.

B. Hachey and C. Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

J.-D. Kim, T. Ohta, and J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

F. Kuhn. 2010. A description language for content zones of german court decisions. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pages 1–7, La Valletta, Malta.

A. Lenci, S. Montemagni, G. Venturi, and M.R. Cutrullà. 2012. Enriching the isst-tanl corpus with semantic frames. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*, page Forthcoming, Istanbul, Turkey.

K. T. Maxwell, J. Oberlander, and V. Lavrenko. 2009. Evaluation of semantic events for legal case retrieval. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2009)*, pages 39–41, Barcelona, Spain.

A. Mazzei, D. P. Radicioni, and R. Brighi. 2009. Nlp–based extraction of modificatory provisions semantics. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 50–57, Barcelona, Spain.

D. McClosky and E. Charniak. 2008. Self–training for biomedical parsing. In *Proceedings of the Association for Computational Linguistics (ACL 2008)*, Columbus, Ohio.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel S., and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP–CoNLL*, pages 915–932.

T. Ohta, Y. Tateisi, and J.-D. Kim. 2002. Genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, CA.

R. Mochales Palau and M.F. Moens. 2011. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 98–107, Pittsburgh, Pennsylvania.

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. available at https://framenet.icsi.berkeley.edu/fndrupal/the_book.

P. L. Spinosa, G. Giardiello, M. Cherubini, S. Marchi, G. Venturi, and S. Montemagni. 2009. Nlp–based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 40–49, Barcelona, Spain.

G. Venturi. 2011a. *Lingua e diritto: una prospettiva linguistico–computazionale*. Ph.D. Thesis, Università di Torino.

G. Venturi. 2011b. Semantic annotation of italian legal texts: a framenet–based approach. In K. Ohara and K. Nikiforidou, editors, *Constructions and Frames 3:1*, pages 46–79. John benjamins publishing company edition.

V. Walker, N. Carie, C. DeWitt, and E. Lesh. 2011. A framework for the extraction and modeling of fact–finding reasoning from legal decisions: Lessons from the vaccine/injury project corpus. *Artificial Intelligence and Law*, 19(4):291–331.

S. Walter, 2009. *Definition extraction from court decisions using computational linguistic technology*, pages 183–224. Berlin–New York: Mouton de Gruyter.

A. Wyner and W. Peters. 2010a. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In *Proceedings of the Legal Knowledge and*

Information Systems Conference (JURIX 2010), pages 127–136, Liverpool, United Kingdom.

A. Wyner and W. Peters. 2010b. Towards annotating and extracting textual legal case factors. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pages 36–45, La Valletta, Malta.

A. Wyner and W. Peters. 2011. On rule extraction from regulations. In *Proceedings of the 24th International Conference on Legal Knowledge and Information Systems (JURIX 2011)*, University of Vienna.

A. Wyner. 2010. Towards annotating and extracting textual legal case elements. In *Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2010)*, pages 9–18, Fiesole, Italy.

Using legal ontology to improve classification in the Eunomos legal document and knowledge management system

Guido Boella[1], **Luigi Di Caro**[1], **Llio Humphreys**[1], **Livio Robaldo**[1]

Dipartimento di Informatica, Università di Torino
Corso Svizzera, 195, 10149, Torino
{boella,dicaro,humphreys,robaldo}@di.unito.it

Abstract

We focus on the classification of descriptions of legal obligations in the Legal Taxonomy Syllabus. We compare the results of classification using increasing levels of semantic information. Firstly, we use the text of the concept description, analysed via the TULE syntactic parser, to disambiguate syntactically and select informative nouns. Secondly, we add as additional features for the classifier the concepts (via their ontological ID) which have been semi-automatically linked to the text by knowledge engineers in order to disambiguate the meaning of relevant phrases which are associated to concepts in the ontology. Thirdly, we consider concepts related to the prescriptions by relations such as deontological clauase and sanction.

Keywords: Legal Text Classification, Legal Ontologies, Information Retrieval

1. Research question

The right to view legislation online is becoming accepted in most countries these days. Legislative XML as a standardised way to structure and cross-reference legislation is one of the greatest achievements of Legal Informatics. The most interesting research in this field is now semantic analysis of content and reuse of existing knowledge to address the problem of the so-called 'resource consumption bottleneck' - that the cost of building semantic resources prohibits development beyond proof of viability. In (Boella et al., 2011), we investigated the plausibility of the best performing method of classification in general on legislative text, improving the performance of the classifier by adding the TULE parser for selecting syntactically significant features and Information Gain for selecting the most informative units. We also used the Confusion Matrix as part of the evaluation process, and envisage its usage as an intuitive and powerful tool for continuous evaluation of classification by non-technical knowledge engineers.

In that paper we posed as future work the use of additional knowledge to improve classification accuracy. These days, more and more legal knowledge is codified in legal ontologies, legal texts are subject to classification and available on the web on institutional portals. In this paper we compare the performance of our classifier on legal text with and without the addition of knowledge built up in ontologies as part of our legal knowledge management system. Our aim is to see to what extent the growth of knowledge can help achieve faster and more efficient analysis and processing of new data.

This research is part of a wider effort to use intelligent technologies for analysing legal and legislative data in the Eunomos legal document and knowledge management system (Boella et al., 2012b). Our overall vision is to to make texts more meaningful and clear for professional users who need to know how law affects their domain of interest. It requires a huge effort to build such knowledge, and to make this effort sustainable and economically profitable, we need to use intelligent technologies as much as possible, from NLP to semantic search, as explained in (Boella et al., 2012a).

The paper is organised as follows. Section 2. describes the Eunomos legal document management system, the ontology it is based on and how it is used to define prescriptions. Section 3. we describe the classification algorithms used and the results of our experimentation on comparing the added value of using additional knowledge. Related work, future work and conclusions end the paper.

2. Description of Eunomos

2.1. A legal knowledge management system

In (Boella et al., 2012b) we introduced the Eunomos software, which is being developed in the context of the ICT4LAW[1] project to address these needs, and compared the product with other systems in the field. Eunomos is an advanced legal document management system based on legislative XML representation of laws which are retrieved automatically from institutional legislative portals, and incorporates a tool for building legal ontologies called Legal Taxonomy Syllabus (Ajani et al., 2007). It provides a powerful knowledge base for analysing laws and keeping up to date with legal changes, although it also requires knowledge engineers to manage the information. Eunomos can be regarded as a desktop where several tools are at the disposal of the legal knowledge engineer to assist with categorising and annotating laws and building ontologies. The system was designed to have the following features:

- A large database of laws (about 70,000 Italian national laws in the current demo) maintained in XML format.
- Automatic downloads of laws from institutional legal portals via dedicated spiders. Currently the software harvests the Italian national portal http://www.normattiva.it including over 50,000

[1]ICT4LAW: "ICT Converging on Law: Next Generation Services for Citizens, Enterprises, Public Administration and Policymakers" funded by Regione Piemonte 2008-2013, call Converging Technologies 2007, website: http://www.ict4law.org

laws, the portal Arianna of Regione Piemonte http://arianna.consiglioregionale.piemonte.it/ and a portal of regulations from the Italian Ministry of Economy.

- The conversion of laws into legislative XML if they are in pure textual format[2] in accordance with the NormeInRete (NIR) standard for Italian laws.
- Automated parsing of legal references using the URN format of NIR.[3] This enables references to be transformed into hypertext links to the relevant legislation, thus facilitating automated linking and user navigation.
- Semi-automated classification of laws at the level of paragraphs or articles according to domains specified by the expert user.
- An alert messaging system, using URN references and semantic similarity tools, that informs users of new laws downloaded into the database and suggests which existing laws could be affected by the new legislation.
- The facility to link concepts from the Legal Taxonomy Syllabus ontology via URN to legal definitions within relevant legislation.

2.2. The Legal Taxonomy Syllabus ontology

The Legal Taxonomy Syllabus ontology on which Eunomos was built was designed to be multilevel from the start. The ontology was originally modelled on European Consumer Law, where terms can mean different things in different languages, domains, and jurisdictions e.g. European versus national legislation. As such, the main assumptions of the Legal Taxonomy Syllabus ontology come from studies in comparative law (Rossi and Vogel, 2004) and ontologies engineering (Klein, 2001). Making a clear distinction among terms and their interlingual acceptions (or *axes*) is a standard way to properly manage large multilingual lexical databases (Sérasset, 1994; Lyding et al., 2006). This approach is also well suited to complex interlinking domains.
The primary focus in the design of the Legal Taxonomy Syllabus lightweight ontology and its integration within the Eunomos system is on enabling connections to be made between relevant knowledge to help contextualise and deepen user understanding of legal terminology (rather than reasoning by intelligent systems). The ontology is hierarchical and allows hyperonomy/meronomy/synonomy relations to be made, and the conceptual tree allows users to easily navigate to the information they require. There are also connections between terms and phrases in legislative sources with definition of associated concepts. When users pass their mouse over terms held in the ontology, popup previews appear with ontological definitions and hyperlinks to further information. The user interface has some similarity to wiki tools. Such enrichment is based on NLP technologies to support manual annotation (Boella et al., 2012a).
Each concept in the terminology has the following fields: language, jurisdiction, domain, description in natural language, notes and links to related concepts, and references to relevant articles. The descriptions in natural language are made by legal experts who seek to explain but do not over-simplify legal issues. The notes field carries information about court decisions, scholarly interpretations or other information of interest. References to relevant articles are made via legislative XML in-document hyperlinks. Each piece of legislation in the Eunomos database is stored in accordance with the Norme in Rete (NIR) legislative XML standard using the ITTIG CNR parser[4]. Each piece of legislation, document, article and paragraph is assigned an XML tag with an Unique Reference Number (URN), which makes it easy to link to the source at any level of granularity.

2.3. The prescription ontology

The Eunomos ontology includes an extended structure for prescriptions, which we described in (Boella et al., 2012c) as individual legal obligations derived from legislation. The prescriptions ontology addresses the problem that it is difficult to fully understand norms directly from legislation, because different aspects of norms are distributed in different sections. The prescriptions ontology can be used by financial institutions (or any other organisation subject to regulatory compliance), to monitor the legality of their business processes.
Each prescription is defined as a concept which is necessarily connected by relations to other concepts as follows:

Deontic clause: the type of prescription: obligation, prohibition, permission, exception.

Active role: the addressee of the norm (e.g., director, employee).

Passive role: the beneficiary of the norm (e.g. customer).

Crime: the type of crime resulting from violation of the prescription, often defined in other legislation such as the Penal Code.

Sanction: a concept describing the sanction resulting from the violation.

The structuring of prescriptions in terms of concepts and relations enables the user to make fine-tuned searches such as 'List the prescriptions for which the director concept has the active role', a most useful feature for the compliance officer.
Figure 1 illustrates a prescription together with its links to other concepts.

[2]The Arianna portal already exports documents to NIR XML format. The conversion in the current version of the software is done using the XMLeges Marker tool developed by Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG) of Florence (http://www.xmleges.org).

[3]This is done using the XML Leges Linker tool developed by ITTIG.

[4]The conversion in the current version of the software is done using the XMLeges Marker tool developed by Istituto di Teoria e Tecniche dell' Informazione Giuridica (ITTIG) of Florence (http://www.xmleges.org).

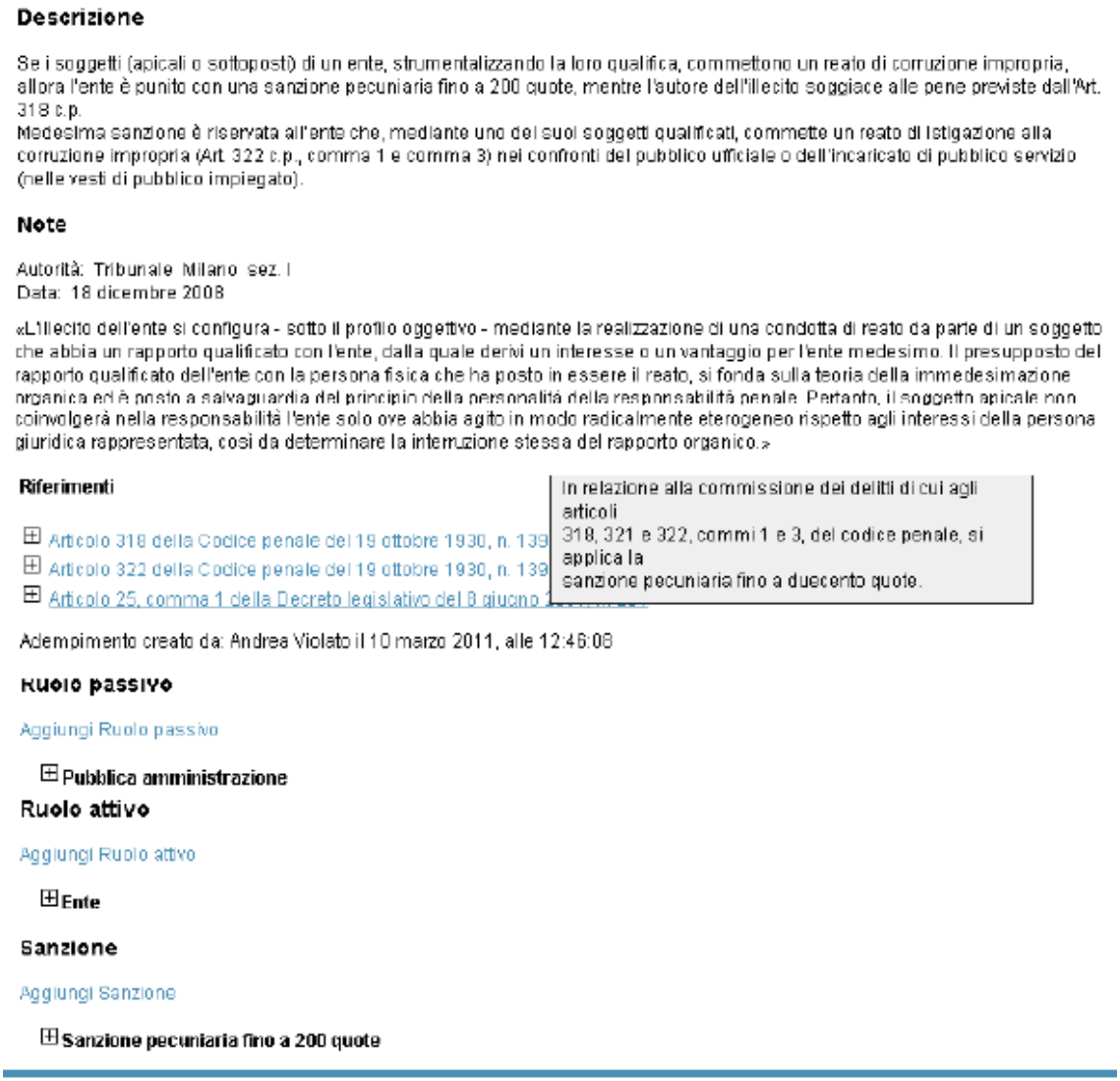


Figure 1: The description of a prescription with the related concepts

3. Eunomos and the classification of prescriptions

3.1. Research task

For this research, we used the descriptions part of the prescriptions as input data to the classifier. We selected these data because a) the data are already labelled with topics (i.e., classes, categories) and b) the classification is at the level of granularity required by professionals requiring knowledge of the law to carry out their work. The data is interesting because it is legal, but not legislative text, and thus might be more representative of the kind of text used in legal correspondence.

The data used for classification contains 156 legal texts organized in 15 classes, each one belonging to a different topic. Since statistical methods requires that the classes should be sufficiently large, we filtered out some with low cardinality, building three datasets with different degrees of filtering (namely, S_1, S_2, and S_3), see Table 1. One class was removed from all the datasets since it contained general-purpose documents and did not reflect a specific class/topic. Note that dataset S_3 preserves more than 70% of the original data (i.e. 110 documents out of 156), although it contains only 5 out of 15 total classes.

In addition to the standard bag-of-words approach (where each text is represented as an unordered collection of words), we also wanted to test whether it is worth using the ontological concepts contained within the legal texts as additional features. Therefore, we added the concept ID numbers of words within the description linked to a concept in the ontology. Finally, we also used the concept IDs of concepts linked to the prescription via relations such as deontological clause, crime, sanction etc. In table 2, the suffix $+CONC$ applied to the name of the datasets indicates the use of only the concept names while $+CONC + REL$ refers to the use of ontological concepts and relationships as well.

3.2. Advanced pre-processing

The process of transforming text into vectors requires selection of suitable terms, and use of a weighting function as part of the frequency calculations. We used the *Term Frequency-Inverse Document Frequency* (TF-IDF) weight-

Class	Description	Cardinality	S_1	S_2	S_3
C_1	Risks evaluation	11	x	x	x
C_2	Contracts	6	x	x	
C_3	Management of emergencies	9	x	x	x
C_4	Controls	5	x	x	
C_5	Information	7	x	x	
C_6	Formation and updating	7	x	x	
C_7	Public health surveillance	4	x		
C_8	Periodic meetings	4	x		
C_9	Communications	5	x	x	
C_{10}	Proscriptions	1	x		
C_{11}	Work environments	38	x	x	x
C_{12}	Work equipments and devices for personal protection	43	x	x	x
C_{13}	Signals for security and health on work	1	x		
C_{14}	Devices with video display terminals	9	x	x	x
C_{15}	General obligations	6			
-	-	156	150	140	110

Table 1: The data used for classification contains 156 legal texts organized in 15 classes, each one belonging to a different topic. Since statistical methods needs the classes to be sufficiently large, we filtered out some with low cardinality, building three datasets with different degree of filtering (namely, S_1, S_2, and S_3). Class $C15$ has been removed from all the datasets since it contains general-purpose documents and thus it does not reflect a specific class/topic. Note that dataset S_3 preserves more than the 70% of the original data (i.e., 110 documents out of 156), although it contains only 5 out of 15 total classes.

ing function as proposed in (Salton and Buckley, 1988), that takes into account both the frequency of a term in a text and how characteristic it is of text belonging to a particular class. There are pre-processing steps that can be carried out on the selection and transformation of terms, which have been shown to be more effective than a simple bag-of-words approach. A commonly-accepted technique is to use a stopword list to remove uninformative terms and morphological transformation of remaining terms to to their lexical roots (i.e., the lemmas). The aim of these procedures is to eliminate noise while reducing redundant linguistic variability. Typically only nouns are then considered as informative features. The accuracy of the classification methods is highly dependent on the quality of these procedures. Our approach differs from standard practice of simply using lists of stopwords and external resources such as WordNet (Miller, 1995) to extract the lemmas, in that we use a dependency parser for Italian called TULE (Lesmo, 2009) that performs a deep analysis over the syntactic structure of the sentences and allows a direct selection of the informative units, i.e., the lemmatized nouns. Moreover, instead of only looking for nouns, we made some further steps. Firstly, we included verbs as factors since they may have some discriminatory power among the classes. Secondly, leveraging the parse trees, we extracted syntactic chunks rather than single word units This approach may increase the system complexity as a whole, but it is a better solution than the standard practice of extracting nouns with top-domain ontologies that are unable to recognise and lemmatize many legal domain-specific terms. Finally, and most importantly, we extracted references to ontological concepts and relationships associated to the input texts, and used these as further additional features for classification. The results in Table 2 shows the improvements gained by the use of this strategy.

3.3. Algorithms and tools

Although there are plenty of algorithms for text classification, we used the well-known Support Vector Machines (SVM) for this task, since it frequently achieves state-of-the-art accuracy levels (Cortes and Vapnik, 1995; Joachims, 1998). This algorithm makes use of vectorial representations of text documents (Salton et al., 1975) and works by calculating the hyperplane having the maximun distance with respect to the nearest data examples. More in detail, we used the Sequential Minimal Optimization algorithm (SMO) (Platt and others, 1999) with a polynomial kernel. The vectorial representation of textual data is particularly complex as it usually contains thousands of features. One property of SVM is its ability to learn from cases with high dimensionality of the feature space (Joachims, 1998), so it fits well with text.

The association between text and a category label was fed to an external application based on the WEKA toolkit (Hall et al., 2009) and incorporated in Eunomos, creating a model that can be used to classify new laws inserted on a daily basis into the database by web spiders or users. The WEKA toolkit (Hall et al., 2009) was used as a framework for the experiments because it supports several algorithms and validation schemes allowing an efficient and centralized way to conduct experiments and evaluate the results.

3.4. Data complexity and evaluation

The evaluation of a classification task can range from very poor to excellent depending on the data. A simple way to estimate the complexity of the input is to compute the separation and compactness of the classes. The *Separability Index* (SI) (Greene, 2001) measures the average num-

ber of documents in a dataset that have a nearest neighbour within the same class, where the nearest neighbour is calculated using Cosine Similarity. Tests on the whole dataset revealed an SI of 66.66%, which indicates a high overlap among the labelled classes. Table 2 shows the SI values for all the three datasets.

The SVM classifier achieves an accuracy of 92.72% when trained with the n-folds cross validation scheme (Kohavi, 1995) on dataset $S_3+CONC+REL$ (using n = 10, which is a common practice in the literature). As shown in Table 2, the classifier achieves lower accuracy levels with datasets S_1 and S_2, though it was already expected from their low SI values. Nevertheless, it is interesting to see that classification on dataset S_1 is still acceptable in terms of accuracy despite its very low SI. This is due to the fact that, although there is a large overlap between the dictionaries used in different classes, there are some terms that characterize them properly. Note that in all cases, the use of ontological factors improves the accuracy of the classification procedure.

The remainder of this paper is concerned with experiments carried out on dataset $S_3+CONC+REL$. Table 3 summarizes the results of the classification in terms of *precision*, *recall*, and *F-measure* for each class in $S_3 + CONC + REL$. While recall indicates how well the classifier recognizes all the documents belonging to one class *C*, precision shows how rarely it classifies as *C* documents which belong to other classes. F-Measure is then a way of merging both indicators into one general measure, calculated as the harmonic mean of precision and recall (Rijsbergen, 1979). As can be noticed in Table 3, the classes with the lower accuracy levels (i.e., C_3 and C_{14}) are the ones with less documents associated with them.

A simple scheme called *confusion matrix* is typically used in supervised learning to see where the classifier is confusing some classes. According to this scheme, each column of the matrix represents the documents in a predicted class, while each row represents the documents in an actual class. The confusion matrix of our classification results for dataset $S_3 + onto$ is shown in Table 4. As can be seen, three documents of class C_{14} are misclassified as C_{12}. This is reasonable due to the scarce quantity of data contained in that class. It should be noted also that classes C_{12} and C_{14} overlap significantly in terms of meaning since they both talk about devices (*Work equipments and devices for personal protection* and *Devices with video display terminals* respectively).

3.5. Discussion

The identification of the topics covered by legal documents is an important task, as it can be used to send targeted email notification to users who are interested in particular domains.

This section has concentrated on the use of ontological information associated to legal text as additional knowledge to improve the performance of a classifier for determining their areas of interest, or topics.

As a first step, we only used the concept names as additional features for the classifier. Then, we tested the integration of a deeper level of semantics coming from the existing relationships between concepts and documents.

The SVM-based classifier achieved high accuracy improvements when trained with such additional knowledge. We have shown how these information are of great help for supervised learning techniques given by their nature of being manually-annotated, i.e., they unconditionally represent highly-informative units (as opposed to automatic approaches like Information Gain for feature selection, as we have previously done in (Boella et al., 2011)).

4. Related work

Concerning text classification techniques in general, there are many algorithms for this task, though Support Vector Machines have repeatedly been shown to be better than Naive Bayes Classifier, Decision Trees, and others (Cortes and Vapnik, 1995; Joachims, 1998).

Concerning text classification for legal text, it is instructive to refer to de Maat et al. (de Maat et al., 2010)'s comparison of machine learning versus knowledge engineering in classification of legal sentences, since Eunomos uses similar techniques. de Maat et al. (de Maat et al., 2010) use a set of rules to find patterns suggestive of a particular class, while we look for patterns containing references to find norms that have been already classified, thus providing a clue as to the classification of new norms. The conclusion of de Maat et al. (de Maat et al., 2010)'s research (ibid, page 16) was that "a pattern based classifier is considered to be more robust in the categorization of legal documents at a sentence level". However, their classification task was quite different since that research was concerned with classifying the type of norms as delegations, penalizations, etc., while we categorize norms as belonging to different topic areas. The author (ibid. page 14) noted that SVMs were better than patterns at categorisation where word order was less restricted. Biagioli et al. (Biagioli et al., 2005) achieved an accuracy of 92% in the task of classifying 582 paragraphs from Italian laws into ten different categories using Multiclass Support Vector Machines. However, they do not classify the text on the basis of their subject matter. Their categories are high-level meta-classes such as 'Prohibition Action', 'Obligation Addressee', 'Substitution', and so on.

Concerning adding related data to improve classification, Sriram et al. (Sriram et al., 2010) and Cataldi et al. (Cataldi et al., 2010) augmented short text from Twitter[5] messages fed into the classifier with additional data taken from the authors' user profile. Cataldi et al. (Cataldi et al., 2009), meanwhile, identified key terms within the text, and added additional related terms based on text corpora analysis.

5. Future work

This research has shown that the addition of ontological definitions and relations as factors can vastly improve the performance of an SVM classifier. The immediate next step in our future work will be to investigate whether the addition of references to legislative articles - by including the URNs or the articles themselves - can improve the results further.

We are also interested in exploring the potential to use relations within the Prescriptions ontology to create other useful classifications beside the legal domain. Given the data,

[5] http://www.twitter.com

Dataset	Separability Index (SI)	Accuracy
S_1	66.66%	71.33%
$S_1 + CONC$	71.34%	**74.66%**
$S_1 + CONC + REL$	72.31%	**78.00%**
S_2	69.28%	74.28%
$S_2 + CONC$	73.57%	**77.85%**
$S_2 + CONC + REL$	74.07%	**82.96%**
S_3	83.63%	89.09%
$S_3 + CONC$	88.18%	**90.00%**
$S_3 + CONC + REL$	91.09%	**92.72%**

Table 2: Separability Index (SI) and accuracy values computed on the three datasets S_1, S_2, and S_3, using the 10-fold cross validation scheme. The accuracy is calculated as the percentage of correctly classified documents on the total. The suffix $+CONC$ applied to the name of the datasets indicates the use of the ontological concepts as additional features while suffix $+CONC + REL$ refers to the use of both concepts and relationships. Note that in all cases, there is an improvement of the accuracy of the classification task.

Class	Cardinality	Recall	Precision	F-Measure
C_1	11	72.7%	100%	84.2%
C_3	9	77.8%	77.8%	77.8%
C_{11}	38	100%	100%	100%
C_{12}	43	100%	87.8%	93.5%
C_{14}	9	66.7%	100%	80.0%
Weighted average	110	92.7%	93.4%	92.4%

Table 3: Classification results using Support Vector Machines (SVM) on dataset $S_3 + CONC + REL$, with the 10-fold cross validation scheme. The average is weighted according to the number of classified documents.

we can build several classifiers, one for each type of information we want to deduce. For instance, given a piece of text, the classifier might be able to suggest the most likely crime and sanction. While legislative text is highly referential, so that norms often specify explicitly which article defines the sanction applies, this is not true of other legal text such as legal correspondence. It should thus be very interesting to the legal profession and compliance industry to see whether such predictions can be made on descriptions of legal obligations.

In view of Eunomos's potential to cater for multilingual and multilevel legal research, we are also interested in investigating whether labelled data from a national domain can be used to classify norms in European legislation or vice versa. Using a classifier trained on material in a different jurisdiction would require a method to map and merge different domain ontologies. To this end, we are interested in the work of Cheng et al. (Cheng et al., 2008b) and (Cheng et al., 2008a), where the authors presented a technique to map concept taxonomies based on textual regulations relying on standard similarity measures such as Cosine Similarity, Jaccard Coefficient (Jaccard, 1901), and context-adapted Market Basket Analysis (Hastie et al., 2005).

Another area for future work will be semi-automated methods for ontology building.

6. Conclusions

The identification of the topics covered by legal documents is an important task, as it can be used to send targeted email notification to users who are interested in particular domains. (Lenci et al., 2007) argued that 'Technologies in the area of knowledge management and information access are confronted with a typical acquisition paradox. As knowledge is mostly conveyed through text, content access requires understanding the linguistic structures representing content in text at a level of considerable detail. In turn, processing linguistic structures at the depth needed for content understanding presupposes that a considerable amount of domain knowledge is already in place.'

In this paper we have made some steps to try and resolve this paradox. We have shown that the use of ontological definitions and relations as factors can vastly improve the performance of an SVM classifier. First, we preprocessed the text via the syntactic parser TULE to select nouns after disambiguation; second, we added associated concepts which were semi-automatically linked by knowledge engineers to the texts; and third, we considered also concepts related to the prescriptions by their specific relations such as active and passive role, crime, sanction, etc. Our SVM-based classifier achieved high accuracy improvements when trained with such additional knowledge. We have shown how these information are of great help for supervised learning techniques given by their nature of being manually-annotated. Our future work will look at improving the accuracy by augmenting the text to be classified with related information.

$\downarrow$ classified as $\rightarrow$	C_1	C_3	C_{11}	C_{12}	C_{14}
C_1	8	2		1	
C_3		7		2	
C_{11}			38		
C_{12}				43	
C_{14}				3	6

Table 4: Confusion Matrix resulting from the SVM classifier trained with 10-fold cross validation, on dataset $S_3+CONC+REL$. The top-left to bottom-right diagonal shows the correct classifications.

7. References

Gianmaria Ajani, Leonardo Lesmo, Guido Boella, Alessandro Mazzei, and Piercarlo Rossi. 2007. Terminological and ontological analysis of european directives: multilinguism in law. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law: ICAIL*, pages 43–48. ACM.

Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. 2005. Automatic semantics extraction in law documents. In *Proceedings of The Tenth International Conference on Artificial Intelligence and Law: ICAIL*, pages 133–140. ACM.

Guido Boella, Luigi di Caro, and Llio Humphreys. 2011. Using classication to support legal knowledge engineers in the eunomos legal document management system. In *Fifth International Workshop on Juris- informatics(JURISIN)*.

Guido Boella, Luigi di Caro, Llio Humphreys, Livio Robaldo, and Leendert van der Torre. 2012a. Nlp challenges for eunomos a tool to build and manage legal knowledge. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*.

Guido Boella, Llio Humphreys, Marco Martin, Piercarlo Rossi, and Leendert van der Torre. 2012b. Eunomos, a legal document and knowledge management system to build legal services. In *Proceedings of AI Approaches to the Complexity of Legal Systems Workshop (AICOL)*, Berlin. Springer.

Guido Boella, Marco Martin, Piercarlo Rossi, Leendert van der Torre, and Andrea Violato. 2012c. Eunomos, a legal document and knowledge management system for regulatory compliance. In *Proceedings of Information Systems: a crossroads for Organization, Management, Accounting and Engineering (ITAIS) Conference*, Berlin. Springer.

M. Cataldi, C. Schifanella, K.S. Candan, M.L. Sapino, and L. Di Caro. 2009. Cosena: a context-based search and navigation system. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, page 33. ACM.

M. Cataldi, L. Di Caro, and C. Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, pages 1–10. ACM.

Chin Pang Cheng, Gloria T. Lau, Kincho H. Law, Jiayi Pan, and Albert Jones. 2008a. Regulation retrieval using industry specific taxonomies. *Artif. Intell. Law*, 16:277–303, September.

Chin Pang Cheng, Jiayi Pan, Gloria T. Lau, Kincho H. Law, and Albert Jones. 2008b. Relating taxonomies with regulations. In *Proceedings of the 2008 international conference on Digital government research*, pages 34–43.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Emile de Maat, Kai Krabben, and Radboud Winkels. 2010. Machine learning versus knowledge based classification of legal texts. In *Proceedings of Legal Knowledge and Information Systems Conference: JURIX 2010*, pages 87–96. IOS Press.

J. Greene. 2001. Feature subset selection using thornton's separability index and its applicability to a number of sparse proximity-based classifiers. In *Proceedings of Annual Symposium of the Pattern Recognition Association of South Africa*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

P. Jaccard. 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142.

M. Klein. 2001. Combining and relating ontologies: an analysis of problems and solutions. In *Workshop on Ontologies and Information Sharing at IJCAI'01*.

R. Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence*, volume 14, pages 1137–1145.

A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi. 2007. NLP-based ontology learning from legal texts. a case study. *Proceedings of LOAIT 07*, page 113.

L. Lesmo. 2009. The Turin University Parser at Evalita 2009. *Proceedings of EVALITA*, 9.

V. Lyding, Elena Chiocchetti, G. Sérasset, and F. Brunet-Manquat. 2006. The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated. In *Proc. of the Wokshop on Multilingual Lan-*

guage Resources and Interoperability at ACL'06, pages 25–31.

G.A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

J. Platt et al. 1999. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*, 208:98–112.

C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

P. Rossi and C. Vogel. 2004. Terms and concepts; towards a syllabus for european private law. *European Review of Private Law (ERPL)*, 12(2):293–300.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November.

G. Sérasset. 1994. Interlingual lexical organization for multilingual lexical databases in NADIA. In *Proceedings of the 15th conference on Computational linguistics (COLING)*, pages 278–282.

B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pages 841–842. ACM.

JurWordNet and FrameNet Approaches to Meaning Representation: a Legal Case Study

Antonio Lazari*, Mª Ángeles Zarco-Tejada**
*Scuola Superiore Sant'Anna
Piazza Martiri delle Libertà, 33, 56127 Pisa-Italy
**Universidad de Cádiz
Avda. Gómez Ulla, s/n 11003 –Cádiz - Spain
*a.lazari@sssup.it **angeles.zarco@uca.es

Abstract

This paper describes JurWordNet, FrameNet and LOIS approaches towards meaning representation regarding the concept 'State Liability' from a cross-linguistic and comparative perspective. Our starting point has been the lexical and conceptual mismatching of legal terms that the process of harmonization in the European Union has manifested. Our study analyzes such concept in Italian, Spanish, French and English and shows how a deeper sub-language based representation of meaning is needed to account for such phenomena. We examine the most important computational- lexical models in an attempt to identify the most suitable and appropriate approach towards lexical-conceptual mismatching of the concept 'State liability' in the European legal tradition. Our proposal shows a formalization of the concept in the four systems mentioned and uses semantic features to represent lexical mismatching and cultural differences. With this study we show in a systematic way the differences in legal tradition and the reasons for divergence in the judicial use of related concepts.

Keywords: conceptual mismatching, legal language, State Liability, semantic features

1. Introduction

The subject of this paper describes part of a project that deals with law, language and meaning representation. We explain and analyze how computational models face the lexical and conceptual mismatching of legal terms that the process of harmonization in the European Union has made evident. As a case study, we consider the lexicalization of the concept of 'state liability' in Spanish, English, French and Italian and examine the computational models JurWordNet (Gangemi et al., 2003), (Sagri, 2004) FrameNet (Baker et al. 2003) and LOIS (Peters et al., 2007) towards meaning representation. This paper describes our first steps towards an evaluation of the most important lexical-conceptual databases in an attempt to identify the most suitable model for the conceptual mismatching of 'state liability'. As our study shows, these systems do not give an account of the conceptual differences through languages and make clear the need for a deeper and more elaborated computational lexical representation.

2. State of the art

The topic of language and law within a computational linguistic environment seems to be a very active one if we consider the last two decades. This is an area that includes linguistic, translation, legal and computational studies. It is, in fact, a very dynamic area where quite a prolific research activity has become apparent with the ever-increasing number of projects and literature dealing with the problem arisen in the EU for the translation of Directives to national legislation. Transposed national legal norms may be very different from other transposed national laws since Union Directives are subject to interpretation. Thus, the multilingual conceptual dictionary Legal Taxonomy Syllabus (LTS) (Ajani et al., 2008) is an attempt to represent legal information in order to solve this problem. The LTS distinguishes terms from concepts in the same language and uses lightweight ontologies with two levels of representation, national legal ontologies and an EU ontology level. Each language ontology is related to the EU ontology via a set of association links.

Even though the LTS does not distinguish between different legal linguistic areas, all the examples shown refer to misalignments (Ajani's term for terminological and conceptual mismatch) occurred in French, German, English and Italian private law. The corpus under analysis does not seem to be of general legal nature but focuses on examples of consumer law.

Following the LTS, Maria Font i Mas (2006) analyzes 17 European Directives, 15 Spanish transposition laws and some Catalan ones and, limiting her analysis to the contract field, she insists on the linguistic problem arisen in the transpositions of European Directives.

On the other hand, the Jur-WordNet project based on the basic underlying principles of WordNet, is conceived as a multi-layered lexical resource and an ontology-based content description model of the legal domain of the ItalEuroWordNet. Based on the DOLCE ontology, allows a more explicit representation of the legal domain and uses Pustejovsky's qualia structure for defining its Top-Ontology level. The project is linked to the Norme in Rete project launched in 1999 with the goal to create a free access portal to normative information. Among other applications, its conceptual knowledge base is aimed at providing information extraction, question answering, automatic tagging, knowledge sharing or norm comparison uses. Thus, the project aims at providing two levels of

information, the semantic relations among legal concepts (synsets) and the ontological relations (Gangemi et al. 2003). JurWordNet is especially interesting and useful for the legal domain since most inter-linguistic mismatching of legal terms has a conceptual nature.

Similarly, one of the most interesting and specific projects to the legal domain is the LOIS project, a multilingual legal lexicon for German, Dutch, English, Portuguese, Czech and Italian (Peters et al., 2007) based on the EuroWordNet framework. This multilingual legal thesaurus enables the comparison of legal systems through the Inter-Lingual_Index, a superset of all concepts from all wordnets. As the authors explain, near-equivalence relations are the most frequent ones in LOIS, since full equivalent concepts are rare, besides hyponym and hyperonym relations.

Another legal-specific FrameNet based approach is the work described by Venturi (2011).

Also related to the very active research field of ontology building, and in contrast to hand-crafted efforts driven by domain experts, the Text-to-Knowledge system (T2K) (Lenci et al., 2009) is an attempt to automatically extract ontological knowledge from Italian legislative texts.

3. Linguistic vs cultural differences

We have chosen the term 'State liability' because it is a key principle in the judicial relations between European Union law and national orders. Our study also deals with the difference between the principle of State liability (State liability for violations of European law norms) and the related criteria 'State responsibility' or 'State accountability'. In the European mainstream it has been argued that, considering the centrality of the figure of the State, a soft-law sanction against the violations of European law would be recommended (Snyder, 1993), (Harlow, 2002). From such point of view a moral and political idea of responsibility has to be placed that, nevertheless, cannot constraint and hold the State responsible through an institute of private law: 'extracontractual liability' (Ponzanelli, 1996).

The fact that with Francovich judgement of 1991 for the first time a judicial sanction for the State failure was established, has provoked a multiform answer in and by national systems. In fact, national models, as Harlow has made clear, may have or not have the key element 'fault'. Besides, in the systems where it exists, such term is colored up by several dyes depending on each particular national culture.

The first element to highlight is the fact that the European law model has to relate to the different national models compulsorily: it is inherently interfacial and it gives rise to a process of cross-fertilization or judicial dialogue. Different from the traditional paradigm, European interfacial norms need a national comparative process that comes out onto several State 'hybrids'.

Thus, first of all, we have to schematically single out the national boxes in which the European paradigm is to be inserted. In France, the administrative State liability system is basically divided in 'responsabilitè pour faute' and 'responsabilitè sans faute'. However, 'faute' has clearly a sense linked to *defaulance* of the bureaucratic apparatus, claiming for a contractual vision of responsibility. It is a *faute anonyme*, where the stress goes onto the function of compensation of the injured, depersonalizing 'faute', that assumes an anthropomorphic dimension in the English model. In England, in fact, the State liability scheme is built up according to an anthropomorphic model based, *in primis*, on the personal responsibility of the public officer according to the Crown Proceeding Acts of 1947 (Kantorowicz, 1957). It is a 'fault' of human shape that draws on Tort figures of private law and on some specific ones such as 'misfeasance in public office' and 'breach of statutory duty'. The Italian State liability model is also based on the general principle of art. 2043 Codice Civile that projects out on the public law screen through the organic theory of the public body. Finally, the Spanish model, unique in the European scene, that does without such element 'fault'.

An analysis of the differences among the liability concept across languages is shown below (Table 1). This model shows a sort of formalization of the differences based on semantic features.

Concept	State liability	
Semantic features	[+fault] [+voluntarily]	
Lexical terms	Breach of statutory duty	Misfeasance of public office
Concept	Responsabilitè	
Semantic features	[+faute administrative]	[-faute administrative]
Lexical terms	Responsabilitè pour faute	Responsabilitè sans faute
Concept	Responsabilità	
Semantic features	[+colpa]	
Lexical terms	Responsabilidad extracontrattuale	
Concept	Responsabilidad	
Semantic features	[+daño]	
Lexical terms	Responsabilidad	

Table 1: feature-based semantic formalization

4. Our proposal

We have checked how the term 'liability' has been represented in FrameNet, JurWordNet and LOIS, and have concluded:

4.1 FrameNet

In FrameNet there is no result for the word 'liability'. Instead, the lexical unit 'responsibility' evokes two different frames, the frame 'responsibility' defined as follows: *An Agent is responsible for having intentionally performed an Act or for being the or a primary instigator behind the Act. There is often a sense of negative evaluation of the Agent for having done so*.

And the frame 'being_obligated' defined as: *Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly.*

From both semantic frames the Frame Elements Duty, Responsible_party, Agent and Act are identified as the core elements to convey meaning.

The lexical unit 'responsibility' is linked to the semantic field of 'intentionality' not only by the explanation supplied as part of the Frame Elements Agent:

Semantic Type: Sentient. The Agent performs the intentional act.

Act:

Semantic Type: State_of_affairs This FE identifies the Act that the Agent performs intentionally

but by the frame-to-frame relations it shows, where the frame 'intentionally_act' is used by the frame 'responsibility': *An abstract frame for acts performed by sentient beings. It exists mostly for FE inheritance*. In it the semantic type for the Frame Element [Agent] is characterized as 'sentient' and defined as *The Agent performs the intentional act*, and the Frame Element Event_description is defined as *This FE gives a description of the Intentionally_act event.*

On the other hand, the lexical unit 'breach' defined as 'break (a law, agreement or code of conduct)' evokes the frame 'compliance' with the following definition: *This frame concerns Acts and State_of_Affairs for which Protagonists are responsible and which either follow or violate some set of rules or Norms.* The meaning related to the fact that in the English tradition 'breach of statutory duty' implies the notion of 'fault' is conveyed in the compliance frame with terms such as 'violate', 'rules' or 'norms' . Even though, the only information concerning the protagonist or how compliance is achieved is coded in the protagonist semantic type, as sentient. Nevertheless, there's no mention to the fact that the act can be realized as a fault or in bad faith, voluntarily or by chance. Even more, the legal concept of 'misfeasance' is not lexicalized in FrameNet. If we analyze the Frame Elements part of the frame Compliance evoked by the Lexical Unit 'breach', - Act, Degree, Manner, Norm, Protagonist-, no semantic role expressing the '(in-)/voluntary', (un)-/consciously' way the act has been committed occurs. The Lexical Units 'compensation' or 'indemnity' do not occur in the FrameNet lexicon.

Regarding the Spanish FrameNet, the equivalent Lexical Unit 'violación' as part of the Spanish 'violación suficientemente caracterizada' corresponding to the English 'breach of statutory duty' is not indexed in the Spanish FrameNet.

4.2 JurWordNet

JurWordNet is organized according to the WordNet principles and, thus, the relational network makes explicit the lexico-semantic relations among concepts. Moreover, not only it makes use of the WordNet relations but it takes advantage of the ones already defined in EuroWordNet, with relations such as ''role_result', 'involed_result' or 'role_agent'. For instance, in our attempt to analyze an equivalent concept to 'breach' in Italian, the corresponding concept 'violazione sufficentemente caracterizata' is considered. The system shows the synset {violazione} linked by hyperonymy to the synset {illecito} and this one is linked by hyponymy to the synsets {illecito contrattuale} and {illecito extracontrattuale} which are, at the same time, linked by an involved_result relation to {responsabilità contrattuale} and {responsabilità extracontrattuale}. The synsets just referred are not defined except for the synset {illecito} as 'atto volontario che viola una norma giuridica ed arreca danno ad un altro soggetto'. No differences between {contrattuale}{extracontrattuale} are mentioned for 'illecito' or 'responsabilità'.

Surprisingly, no relation is established between the synset {responsabilità} and {risarcimento} even though {responsabilità} is defined as 'situazione giuridica in cui un soggetto **deve rispondere** di un obbligo inadempiuto o di un atto illecito'. We claim to set a 'CAUSES' and 'IS_CAUSED_BY' relation between the synsets {responsabilità} and {risarcimento}. Following Vossen (2002) the cause relation is used to link 2ndOrderEntities, which can be either verbs, nouns or adjectives. There is a constraint on the causing event that should be dynamic. According to the tests proposed by the author, in a factitive causation relation, X causes Y to take place or, X has Y as a consequence or, X leads to Y. In our case: {responsabilità} CAUSES {risarcimento}.

In addition, the synset {indennizzo} seems to have no relation to {responsabilità} either. The only hyponym relations are established with {indennizzo assicurativo} and {indennizzo danno} and these ones, at the same time, to {indennizzo} by hyperonymy. No definitions are supplied. We believe there is a conceptual relation between the synsets {responsabilità}, {risarcimento} and {indennizzo} and strongly suggest the need to define and differenciate {risarcimento} from {indennizzo}: as figure 1 shows below, 'indennizzo' is not linked to a tort act and thus it should be related to liability for legal act, whereas 'risarcimento' is related to injury or breach of duty.

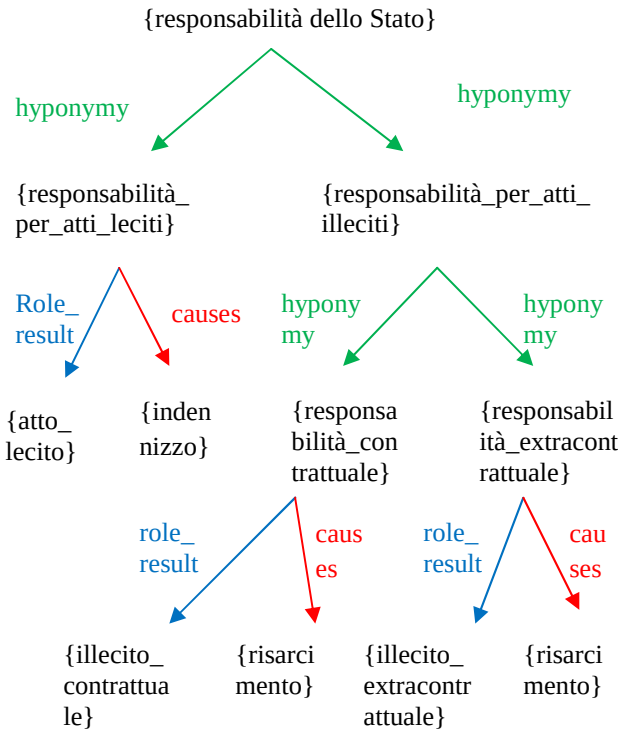


Figure 1: lexical relations with CAUSES

4.3 The LOIS project

The LOIS project, as a multilanguage legal thesaurus, is the best and most appropriate approach towards multilingual conceptual representation of legal terms. This legal knowledge repository, that contains terminology within the domain of consumer law, shows very interesting links for the synset {liability}. Linked by hyperonymy to {obligation}, shows a great variety of hyponymy relations such as {liability in contract}, {liability in tort}, {debt} and {objective liability}. This latter synset describes a related conceptual liability within the State Liability English tradition arising by statute only. We show in table 2 below some of the relations:

Semantic context of liability in LOIS	
Relations	**Synsets**
has_hyperonym	obligation
has_hyponym	absolute liability
has _hyponym	contract liability
has_hyponym	tort liability
has_hyponym	objective liability
has_hyponym	debt
has_hyponym	disadvantage
has_hyponym	obligation to pay interest
has_liability	responsible
involved_patient	person responsible

Table 2: relations in LOIS

4.4 Our model

Besides the fact that the most representative systems do not have all these terms indexed and when they appear they lack definitions or important legal details that should be shown as part of the lexical-semantic relations, there is a common deficiency in all, that is, the difference if any between the concepts of compensation and indemnity and their conceptual relations to liability in all languages under study.

We think JurWordNet and FrameNet could be improved by adding the 'CAUSES' and 'IS_CAUSED_BY' relation for the first one or by expanding the frame being_obligated in FrameNet. The differences between 'contrattuale' / 'extracontrattuale' could be codified as part of the FE Condition, important conceptual difference if we want our system to include 'risarcimento' and 'indennizzo' as part of the FE Consequences.

At the same time, the FE Duty could be expanded via the FE Act, part of the Compliance frame –defined as *This frame concerns Acts and State_of_Affairs for which Protagonists are responsible and which either follow or violate some set of rules or Norms*-, related by the frame-to-frame relation through the obligation_scenario. This way, the FE duty as part of the definition of the frame beign_obligated could be specified as a legal or an illegal act (important difference between atto lecito or atto illecito).

The English system is based on the personalization of the tortfeasor (Agent) more than on compensation. This is shown in the deterrence function that the personal liability principle has in the English tradition (Lazari, 2005). Such an aspect, backbone of the English system, is formalized through the semantic features described in table 3 where the most characteristic features for each system are highlighted in bold. In the English system, liability for torts (strict liability is just a legal niche provided by the legislator sporadically) is predominant. The torts of misfeasance in public office, breach of statutory duty and negligence have all a common feature [+fault], characterized by the anthropomorphic feature [+sentient], [+personalized]. Within these torts the English system rigorously envisages the kind of tort in Misfeasance in public office, thus [+intentional_wrong_doing].

Quite different from the English system, the French one shows a clear tendency to depersonalization of the tortfeasor as part of the idea of faute that is [-sentient] and [-personalized] within the predominant category of responsabilitè pour faute simple. The French model is not completely objective as long as it envisages a faute lourde only for exceptional cases that could be considered close to the depersonalized English misfeasance [±sentient], [±personalized]. On the other hand, the legal feature reponsabilitè sans faute is provided where the accent is placed on the characteristics of the special dogmages and in the absence of any faute administrative. The function of compensation more than deterrence is evident.

The tendency to depersonalizing the administrative attitude is confirmed in the Italian order too. Here we

do not have to distinguish between fault system or no-fault system, but identify the presence of illecito [+wrong_doing]. If the conduct would get criminal nuances [+intention_wrong_doing], the residual feature of danni morali could be generated. Due to this feature of illicità (wrong_doing), the Italian system restricts the compensation function very much. The concept of indennizzo remains just as a residual niche of contractual liability, where there is no fault.

The Spanish scheme is structurally closer to the French system, which is even more compensation-friendly because it does without fault theoretically [-fault], [-personalized], [-wrong_doing]. It is for this reason that it has been related to the institute of compulsory purchase more than to contractual liability. In fact, in the Spanish order *resarciendo por funcionamiento normal o anormal*, the difference between lawful or unlawful act is not distinguished: everything is indemnity.

5. Conclusions and further work

A more precise conceptual knowledge base approaches or more detailed frame base systems allow to establishing conceptual correspondences among terms in different languages or, furthermore, allow to showing evidence of inter-cultural conceptual mismatching of legal terms. As Sagri, M.T, Tiscornia, D. and Bretagna, F. (2004) states "in law we do not speak of the translation of a legislative text but rather of its multilingual versions. The issue […] is crucial in European Community".

It seems clear that we need deeper linguistic devices to represent language conceptual differences. The data under study strongly suggest the need for highly detailed representational meaning useful for, among other things, norm comparison and cross-linguistic judicial dialogue. Our study has focused on the concept 'liability' and 'responsibility' that clearly shows conceptual and lexical mismatching across languages and has made evident the possible devices to account for such phenomena. A more fine-grained representation of meaning as part of the concept-based or frame-based systems could explain the different legal traditions.

Our future work will be twofold. First we will focus on the analysis of semantic features, when representing conceptual mismatching of legal terms, and their relation to legal formal ontologies, and, second, we will follow a sentence-driven approach towards the formalization of cross-linguistic lexical and conceptual mismatching oriented to the identification of cultural convergence and divergence in the European law context.

6. References

Ajani, G.; G. Boella, L. Lesmo, A. Mazzei and P. Rossi. (2007). Multilingual Ontological Analysis of European Directives. In Proceedings of Association for Computational Linguistics, 2007.

Ajani, G., G. Boella, L. Lesmo, A. Mazzei, D.P. Radicioni, and P. Rossi. (2008). Legal Taxonomy Syllabus: Handling Multilevel Legal Ontologies. In *Proceedings of Langtech*, February.

Baker, C. F.; Fillmore, C. J.; Cronin, B. (2003). The Structure of the Framenet Database. In *International Journal of Lexicography* 16.3:281-296.

Font i Mas, M. (2006). Qüestions terminològiques en la transposició de directives sobre dret contractual europeu: la perspectiva de l'ordenament jurídic español. In *Revista de Llengua y Dret*, n. 46.

Gangemi, A.; Sagri, M.T. and Tiscornia, D. (2003). Metadata for content description in Legal information. In *Proceedings of LegOnt Workshop on Legal Ontologies*, ICAIL 2003, Edinburgh.

Harlow, C. (2002). Voices of Difference in a Plural Community". In P. Beaumont, C. Lyons and N. Walker (Eds.), *Convergence and Divergence in European Public Law*, Hart Publishing, pp. 199-204.

Kantorowicz, E. (1957) *The King's Two Bodies*. Princeton, New Jersey: Princeton University Press.

Lazari, A. (2005). Modelli e Paradigm della Responsabilità dello Stato. Torino: Giapichelli.

Lenci, A.; Montemagni, S.; Pirelli, V.; and Venturi, G. (2009). Ontology learning from Italian Legal texts. In *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, Amsterdam, The Netherlands: IOS Press.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D. and Miller, K.J. (1990). Introduction to WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, vol.3 n.4, pp. 235-244.

Peters, W.; Sagri, M.T.; Tiscornia, D. (2007). The Structuring of Legal Knowledge in LOIS. In *Journal of Artificial Intelligence and Law*, vol15, pp.117-135.

Ponzanelli, G. (1996) *La responsabilità civile. Profili di diritto comparato.* Bologna: Il Mulino.

Sagri, M.T.; Tiscornia, D. and Bretagna, F. (2004). Jur-WordNet. In *Proceedings of the Second International WordNet Conference*, pp. 305-310.

Snyder, F. (1993).The effectiveness of European Community Law: Institutions, Processes, Tools and Techniques. In *Modern Law Review*, vol. 56, pp. 19-56.

Tiscornia, D.; Francesconi, E. and Spinosa, P.L. (2007). A Linguistic-ontological Support for Multilingual Legislative Drafting: the DALOS Project. LOAIT 2007: 103-111.

Venturi, G. (2011) Semantic annotation of Italian Legal Texts: a FrameNet-based Approach. In K. Ohara and K. Nikiforidou (Eds.) *Constructions and Frames*, John Benjamins Publishing Company.

Vossen, P. (2002) EuroWordNet General Document, Version 3, Final, in http://www.hum.uva.nl/~ew.

Table 3

Systems	**ENGLISH SYSTEM**					**FRENCH SYSTEM**				
concept	State liability					Responsabilité de l'État				
	Contractual liability	**Extracontractual liability**				Contractual liability	Extracontractual liability			
		Objective liability	**Tort liability**				***Objetive liability***	***Tort liability***		
Lexical terms		Strict liability	***Misfeasance in public office***	***Breach of statutory duty***	***Tort of negligence***	Responsabilité contractuel	**Responsabilité administrative**			
							Responsabilité sans faute	**Responsabilité pour faute**		
								Pour faute simple	**Pour faute lourde**	
Semantic features		[-fault]	***[+fault]***				***[-faute]***	***[+faute]***		
		[-sentient] [-personal.]	***[+sentient] [+personalized]***					***[-sentient] [-personalized]***	***[±sentient] [±personalized]***	
			[+intentional w.]	***[-intention.]***	***[-intentional w.]***			***[-intentional wrong]***	***[+intentional wrong]***	
Lexical relations: causes		indemnity	damages				indemnisation	dommages		
			Punitive	Ordinary						
Functions		compensation	**deterrence**				***Réparation***		dissuasion	

Systems	**ITALIAN SYSTEM**				**SPANISH SYSTEM**			
concept	Responsabilità dello Stato				Responsabilidad del Estado			
	Contractual liability		***Extracontractual liability***		Contractual liability	***Extracontractual liability***		
Lexical terms	Responsabilità contrattuale		***Responsabilità extracontrattuale***		Expropiación forzosa	***Responsabilidad extracontractual***		
	Per atti leciti	Per atti illeciti	***Per atti illeciti***			Por funcionamiento normal	**Por funcionamiento anormal**	
			Danni patrimonali	***Danni morali***				
Semantic features	[-fault] [-person.] [-wron.]	[-fault] [-personalized] [+wrongdoing]	***[-fault] [-personalized] [+wrongdoing]***	***[+fault] [+personalized] [+intentional wr.]***	[-fault] [-personalized] [-wrongdoing]	[-fault] [-personalized] [-wrongdoing]	**[-fault] [-personalized] [-wrongdoing]**	
Lexical relations: causes	indennizzo	***risarcimento***			indemnización	***indemnización***		
Functions	compensation	***deterrence***			compensation	compensation		

A Rule-based Parsing Approach for Detecting Case Law References in Italian Court Decisions

L. Bacci, E. Francesconi, M.T. Sagri

ITTIG-CNR, via de' Barucci 20, Florence (Italy)
{bacci,francesconi,sagri}@ittig.cnr.it

Abstract

In this paper a procedure able to detect legal references in Italian court decisions, providing automatic document hyperlinking is described. It is based on the adoption of a naming convention for case law documents, based on the metadata typically used in citations. The parsing strategy in particular is based on regular expressions, able to extract, from legal citations, the metadata used in the adopted naming convention. In particular the parser is able to implement both the ECLI and the LEX naming conventions for case law material.
Keywords: Legal documents naming convention, Case law textual parsing, Regular expressions.

1. Introduction

The analysis and mark-up of references in legislative and case-law documents represent an important pre-condition to provide Web users with effective support for legal documents browsing and retrieval. They also pave the way to the analysis of citation network able to give an idea of the complexity of a legislative corpus (Mazzega et al., 2009) and for determining the authority of a specific piece of legislation or case-law (Winkels et al., 2011; Kirchberger, 2011).
Such analysis, together with the adoption of a naming convention providing unequivocal identifiers for legal documents, allows us to establish persistent and reliable hyperlinks to legal documents, independently from their physical locations, availability and actual publication. In particular the adoption of a naming convention based on metadata typically used in citations (Francesconi, 2011) allows one to implement automatic document hyperlinking procedures (de Maat et al., 2006; Biagioli et al., 2005) for both legal and non-legal texts (ex. newspaper articles) which might contain legal document citations.
In this paper a procedure able to detect legal references in Italian court decisions, providing automatic document hyperlinking is described.
In particular, in Section 2. an analysis of the textual citations within a document dataset has been carried out; in Section 3. two main naming conventions adopted for describing references are introduced; in Section 4. a case law citations parsing strategy able to detect relevant metadata, used in the introduced naming conventions, is described; in Section 5. the methodology adopted for developing some preliminary parsing tests is shown; finally, in Section 6. some conclusions are reported.

2. Analysis of citations in Italian case law

In supranational contexts like the EU, legal practitioners are required to orchestrate and interpret legal rules to adapt to changing social needs or EU requirements. In a Web context this means accessing relevant cases through hypertext links, within a dense network of references to other pertinent cases, as well as to the set of previous judicial and legislative sources relating to a particular legal issue. For years legal publishers have faced this problem by highlighting or including references in the judicial texts through very expensive editorial activities.
The case law references parsing startegy presented in this work, joint with the adoption of a naming convention for case law, allows us to automate this process and may represent an effective solution for legal document management for both legal publishers and courts. A pre-condition to develop such solution is the analysis of the different typologies of case law references as expressed for Italian court decisions, in order to represent such citation patterns in a formalized grammar.
The analysis of such citation patterns has been carried out in a dataset provided by the Court of Milan. Such analysis showed a great variability in the textual forms of jurisprudential references which are different from the legislative ones. In fact in the last years particular attention has been paid to legislative drafting, through the adoption of specific guidelines, including rules for uniform legislative references wording. Similar attention for jurisprudential references has emerged only recently, but decisions are still drafted without any specific guidelines, in particular as far as references are concerned.
In this analysis, inconsistent and incorrect references have been found, even if the practice of using standards for legal document identification (Spinosa et al., 2012) in the last few years seems to have harmonized legislative references wording. In Table 1 different textual forms of normative references are reported: in particular different modalities to cite the same legislative decree 31 December 1992, n. 546 are shown.

Italian	**English translation**
Art. 15 Decreto Legislativo 9 ottobre 2002, n. 231	Art. 15 Legislative Decree 9 October 2002
decreto legislativo 231 del 2002	legislative decree 231 of 2002
il decreto legislativo 231/2002	legislative decree 231/2002
art. 4 del decreto legislativo n. 231/2002	art. 4 of the legislative decree n. 231/2002

Table 1: Different normative references to an Italian legislative decree

On the other hand the analysis of jurisprudential references revealed more variability. In the following examples this variability is shown: for example in many references docu-

L. Bacci is author of Sections 4 and 5; E. Francesconi is author of Sections 1, 3, 5; M.T. Sagri is author of Section 2.

ment number and date are differently written. Examples of this can be found in Table 2

Cass. n. 2597 del 7/2/2006	Cass. n. 13085 del 2008
Cass. 1197 del 20/1/2006	Cass. 19/10/2007, n. 2200
Cass. SU 4.11.2004 n 21095	Cass. 15712/02, 2059/00.
Cass. 21.12.2002 n. 18223, 3.9.96 n. 8053.	

Table 2: Examples of non-homogeneous judicial document citations

In other cases, however, (see Table 3), precedents are cited in support of a legal argument, but without making explicit reference to the matter (civil or criminal) of the dispute. The drafter assumes that the references are on the same subject of the current act, but this practice might be a problem in the retrieval phase.

Italian	English translation
[civil law] Cass. n. 2597 del 7/2/2006; (cfr Cass. SU 4.11.2004 n. 21095)	Cass. n. 2597 del 7/2/2006 (see Cass. SU 4.11.2004 n. 21095)
[criminal law] Cass. Sez. 5, Sentenza n. 8282 del 2006	Cass. Sec. 5, Judgment n. 8282, 2006

Table 3: References of Supreme Court of Cassation with implicit sections

A different kind of problem is related to the practice of referring to judicial authorities in different ways. In the Table 4, for instance, references to decisions or ordinances of the Supreme Court of Cassation (Italian High Court of last resort) are reported. The Court of Cassation is divided into five civil sections, including labor and tributary sections, as well as seven criminal sections. For particularly important cases or in those with conflicting orientations of the sections, the Supreme Court convenes in united sections. The reported examples underline that there are many different types of references relevant to the Court, or cases in which, by omitting the reference to a section, it is difficult to understand whether the measure belongs to a specific section or to the united sections.

In Table 5 some problematic cases are reported in which the drafter uses non-standard methods of citations.

3. Case law naming conventions

The variability of case law citations, reported in Section 2., might represent a source of problems for the editorial activity of detecting and organising references to legal sources, in order to transform them in effective hyperlinks for Web publishing. In this case, in particular, a further problem can emerge which is due to the potential loss of validity of document hyperlinks, which might require intensive maintenance activities.

In order to provide effective and stable hyperlinks to legal documents, recently a number of naming conventions for legal sources have been proposed that are independent from the availability, publication and physical location of the resources. They are based on formal parameters of citations, basically metadata of an act, usually expressed in textual citations by using natural language. Such parameters can be extracted from such citations by textual parsers able to automatically detect and analyse legal references, and can be serialized according to a particular naming convention, representing the effective instrument to establish stable hyperlinks to the cited acts.

Italian	English translation
References to the Supreme Court of Cassation United Sections	
Corte di Cassazione, SS.UU. civili, ordinanza 26 maggio 2004 n. 10180	Court of Cassation civil UU.SS., order 26 May 2004 n. 10180
Cass., se. un., 26 febbraio 2004, n. 3948	Cass., un. sec., 26 February 2004, n. 3948
Cass. S.U. 12 gennaio 2010, n. 262	Cass. U.S. 12 January 2010, n. 262
Cass. ord. 19958/2005	
con pronuncia a Sezioni unite l'indirizzo stato ribadito nella sentenza n. 8014 del 02/04/2009:	with judgement in United Sections, the orientation has confirmed in the decision n. 8014, dated 04/02/2009:
Cass. S.U. 553/2009	
References to Court of Single Sections	
III Civil Section	
Cass., sec. III, n. 12740 del 2001	
Cass. sec. 3, 22.5.2007 n. 11888;	
Cass. sec. 3, 16236/05;	
VI Criminal Section	
Cass. Sez. 6^ n. 113 del 18.1.1993 rv. 193345 e n. 2667 del 1.10.1993 rv. 196598	Cass. pen., sez. VI, 02/10/2003, n.43492, in Guida al Diritto, 2004, 8, 83.
Implicit References to sections of a court	
Cass. 10.11.97 n. 11047, 12.12.97 n. 12572, 17.9.97 n. 9257).	
References Counstututional Court	
La sentenza n. 274/05 della Corte Costituzionale	Judgment n. 274/05 of Constitutional Court
Corte Cost. 24 novembre 1982, n. 196;	Const. Court. 24 November 1982, n. 196;
Corte cost. n. 52/1984; n. 301/1986; 395/1988, 424/1988; 585/2000	Const. Court. n. 52/1984; n. 301/1986; 395/1988, 424/1988; 585/2000
Corte Cost. 105/72; 94/1977; 16 e 18/1981; 348/1990; 112/93	Const. Court 105/72; 94/1977; 16 and 18/1981; 348/1990; 112/93

Table 4: Examples of non-homogeneous references to the sections of a court

Italian	English translation
sentenza n. 10375/08, pubblicata il 19.5.08, con la quale il Tribunale capitolino	decision n. 10 375/08 published on 19.5.08, with which the "Tribunale capitolino" (the expression means the Court of Rome)
al termine della compiuta istruttoria, con sentenza n. 4977 del 15.9 / 25.11.2006, il Tribunale adito rigettava le domande delle ricorrenti	at the end of the prelimary examination, with judgment n.4977 15.9 4977 / 25.11.2006, the resort Court rejected the plaintiffs' applications
il Tribunale adito, con sentenza n. 122 del 2001	the resort Court, with judgment n. 122, 2001
si richiama la sentenza a S.U. della Suprema Corte n. 1786/2010	it refers the judgment Supreme Court United Sections n. 1786/2010 (referring to the Civil Court of Cassation)

Table 5: Examples of non standard citations pattern.

The LEX initiative (Spinosa et al., 2012) for example provides a naming convention for legislative, case law and administrative documents at international level; on the other hand the ECLI[1] initiative represents a European institutional attemp to provide unequivocal identifiers to EU case law documents and to harmonize their modalities of citations (van Opijnen, 2008; EU, 2011).

In our work to develop a parser able to automatically detect references in Italian case law documents, we have adopted both standards which basically rely on the same set of metadata, here below briefly summed-up.

[1] European Case-Law Identifer

3.1. LEX naming convention

The LEX naming convention is based on a URI syntax, namely a string of characters used to identify a name or a resource on the Internet. In particular a URN syntax is used: it identifies a resource through a name within the urn:lex namespace. Nevertheless an http-based version of the standard has been also defined[2], allowing the LEX naming convention to be compliant to the most recent directions of the Semantic Web, in particular the set of principles and technologies, known as "Linked Data", promoting http-based URI as standard for naming resources.
The LEX naming convention is based on FRBR[3] model developed by IFLA[4]; moreover it follows the specifications given in the CEN Metalex (Boer et al., 2009) initiative which provides the definition of a standard interchange format for sources of law, including recommendations for their naming.
The identifier has a hierarchical structure as follows:

```
"urn:lex:"<NSS>
```

where "urn:lex" is the Namespace, which represents the domain in which the name has validity, as well as NSS is the Namespace Specific String composed as follows:

```
<NSS>::=<jurisdiction>":"<local-name>
```

where:
`<jurisdiction>` is the part providing the identification of the jurisdiction, generally corresponding to the country, where the source of law is issued. It is also possible to represent international organizations (either states or public administrations or private entities);
`<local-name>` is the uniform name of the source of law in the country or jurisdiction where it is issued; its internal structure is common to the already adopted schemas. It is able to represent all the aspects of an intellectual production, as it is a legal document, from its initial idea, through its evolution during the time, to its realisation by different means (paper, digital, etc.).
The `<local-name>` of a source of law is structured according to the FRBR model. As in any intellectual production, 4 fundamental entities (or aspects) can be specified. The first 2 entities reflect its contents:

- Work: identifies a distinct intellectual creation; in our case, it identifies a source of law both in its being (as it has been issued or proposed) and in its becoming (as it is modified over time);
- Expression: identifies a specific intellectual realisation of a work; in our case it identifies every different (original or up-to-date) version of the source of law over time and/or language in which the text is expressed;

while the other 2 entities relate to its form:

- Manifestation: identifies a concrete realisation of an expression; in our case it identifies realizations in different media (printing, digital, etc.), encoding formats (XML, PDF, etc.), or other publishing characteristics;
- Item: identifies a specific copy of a manifestation; in our case it identifies individual physical copies as they are found in particular physical locations.

Citations usually address a resource at Work level, by providing metadata specifications for the unequivocal identification of a source of law.
According to the LEX specifications the structure of the document identifier at work level is made of the four fundamental elements, chosen from those used in citations:

```
<work>::=<authority>":"<measure>":
      "<details>[":"<annex>]*
```

where:
`<authority>` is the issuing or proposing authority of the measure (e.g., State, Ministry, Municipality, Court, etc.);
`<measure>` is the type of the measure both public nature (e.g., constitution, act, treaty, regulation, decree, decision, etc.) as well as private one (e.g., license, agreement, etc);
`<details>` are the terms associated to the measure, typically a date and a number;
`<annex>` is the identifier of the annex, if any (e.g., Annex 1).
The main elements of the national name are generally divided into several elementary components, and, for each, specific rules of representation are established (criteria, modalities, syntax and order)[5]. Examples of `<work>` identifiers as regards legislation and case law:

Act n. 22, 14 May 2006 (IT)
`urn:lex:it:stato:legge:2006-05-14;22`
Legislative Decree of the Ministry of Justice, n. 45, 07 October 1999 (UK)
`urn:lex:uk:ministry.justice:decree:1999-10-07;45`
Act of the Glarus canton n. 963, 15 October 2007 (CH)
`urn:lex:ch;glarus:regiere:erlass:2007-10-15;963`
Decision of the Supreme Court n. 68, 28 September 2001 (ES)
`urn:lex:es:tribunal.supremo:decision:2001-09-28;68`
Bill n. 1762 of the French National Assemblee in the XIII legislature (FR)
`urn:lex:fr:assemblee.nationale:proposition.loi:13.`
`legislature;1762`
Brazilian Constitution, 05 October 1988 (BR)
`urn:lex:br:estado:constituicao:1988-10-05;lex-1`
Free Software Foundation, General Public License, 29 June 2007
`urn:lex:fsf.org:free.software.foundation:general.public.`
`license:2007-06-29;lex-1`
Decision of the Supreme Court n. bc8581, 01 April 2008 (NL)
`urn:lex:nl:hoge.raad:besluit:2008-04-01;bc8581`

The use of citation elements at Work level allows one to construct the LEX identifier of the cited act manually or by software tools, as presented in Section 4., implementing automatic hyperlinking of legal sources on the basis of the textual citations of the acts.

[2] from v.06 of the LEX naming convention `https://datatracker.ietf.org/doc/draft-spinosa-urn-lex/`
[3] Functional Requirements for Bibliographic Records
[4] International Federation of Library Associations and Institutions
[5] For the details regarding each element, see Attachment B of the IETF Internet Draft http://datatracker.ietf.org/doc/draft-spinosa-urn-lex/

3.2. ECLI naming convention

As previously mentioned, the ECLI naming convention is an institutional European initiative for providing case law documents with unequivocal identifiers on the basis of a URI-like syntax. Differently from the LEX naming convention, it is defined within an http-based namespace (`https://e-justice.europa.eu/ecli`), while the name follows an URN-like syntax. ECLI naming convention address only the Work level of the FRBR model (EU, 2011). An ECLI identifier contains the following five components, appearing in the listed order:

(a) the abbreviation "ECLI";

(b) the country code for the country under whose competence the judicial decision is rendered;

  (i) For Member States and candidate countries the codes in the Inter-institutional style guide [6] are used;

  (ii) for other countries ISO 3166 alpha-2 is used;

  (iii) for the European Union the code "EU" is used;

  (iv) for international organizations a code is decided upon by the European Commission, taking into account the codes starting with "X" as already being used by European institutions;

(c) the abbreviation for the court or tribunal (court code);

(d) the year of the decision, which must be written in four digits;

(e) an ordinal number, which must be unique in the sense that there must not be more than one judgment of the same court within the same year with the same ordinal number.

All components are separated by a colon (":") while letters in all of the components must be Latin alphanumerical characters only, written in capitals.
The court code includes from one to seven characters. Such code begins always with a letter and has to be chosen in such a way that it appears logical to people familiar with the organisation of the judiciary of the country concerned. For these reasons an abbreviation of the name of the court or tribunal, including also an indication of the chamber or division within that court or tribunal, is recommended. The court code is established by an ECLI country coordinator.
Examples of citations of decisions represented according to the ECLI standard are reported in Table 6

Case citation	**ECLI representation**
Decision of the Italian Corte Costituzionale 322/2008	ECLI:IT:CC:2008:322
Dutch Hoge Raad decision LJN BC8581 of 01-04-2008	ECLI:NL:HR:2008:BC8581
Belgium Raad van State/Conseil d'État, 185.273 of 9 july 2008	ECLI:BE:RSCE:2008:185273.

Table 6: Citations of decisions represented according to ECLI

[6] `http://publications.europa.eu/code/en/en-370100.htm`

4. Case law parsing approach

In order to implement an automatic hyperlinking of documents towards case law decisions, a parser called *Prudence* has been developed. It aims at identifying the textual references to judicial documents, extract the metadata components which can be identified in textual citations, and serialize them according to either LEX or ECLI naming conventions. Such parser has been recently developed for the Italian language by ITTIG-CNR in the course of a collaboration with the Court of Milan and it estabilishes the foundation for a solid semantic analysis of decisions.
Currently the parser runs as a Java web application within a JBoss environment[7]. Through a web form the user can select and upload a decision of a court in pdf format. The output consists of a list of text fragments in which at least one reference to a judicial document is found, including references to other decisions, orders or decrees of a court. Moreover, the details of the cited documents, like the enacting authority, date and numbers, are identified and marked up.
A similar problem aimed at detecting normative references in legislative texts was faced in (de Maat et al., 2006; Francesconi, 2006). However the nature of case law references in decisions is quite different from the normative references contained in legislative texts. The text of decisions is typically more verbose, less formally structured and the details of a reference are often mixed with sparse text or expressions leaving important details as implicit, causing ambiguity not easy to solve. On the other hand, the number of all the possible textual expressions of a judicial reference, though high, doesn't justify the effort of a machine learning approach.
The parser actually implements four finite state automata, based on regular expressions and exclusive start conditions, running in turn and providing an incremental mark-up of the document. This step-by-step approach significantly reduces the ambiguous situations. The lexical analyzer has been generated using JFlex[8]. In Fig. 1 the main phases of the *Prudence* parsing strategy are sketched.
At first, the textual content is extracted from the pdf file specified as input, then a structural analysis of the text is performed, aiming at identifying the single pages and the single sentences. The sentences are eventually used as context of the included references. In this phase the parser operates a reduction of the textual noise like page numbering, headers, etc. Afterwards, all the possible enacting authority, including courts, sections, councils, judges and geographical details, are identified. Then dates and numbers (particularly the file numbering of the court) found in the text are marked up. Finally, exploiting the actual mark-up and the analysis of more expressions relative to the different types of judicial documents (decisions, orders and decrees), the parser tries to build up a valid judicial reference, including partial and multiple ones. The reference is eventually saved with his context (the sentence or the textual fragment where the reference is found).
The details of the reference are at this point filled with raw

[7] http://www.jboss.org
[8] http://www.jflex.de

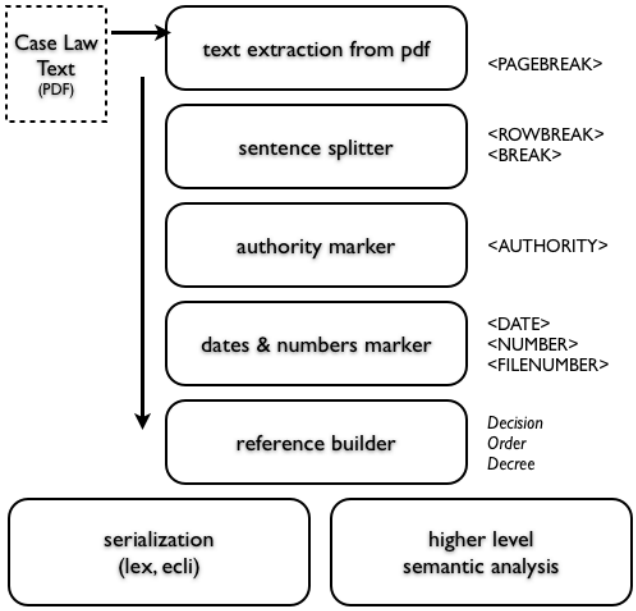


Figure 1: Phases of the *Prudence* parsing startegy

text. A process of normalization can now be performed in order to fit standard formats like LEX or ECLI syntax, or to adapt them to the specification of a relational database for document management and information retrieval. Moreover, the modular policy of *Prudence* makes it possible to connect with a higher level semantic analysis, exploiting the mark-up of both the structure and content, the context and the raw data details of the references.
A note-worthy feature of the parser consists of the special treatment reserved for the details of the input document, usually found at the beginning of the text. They are identified and conveyed as part of the output and can be easily used, for example, to archive in a digital format the document and make reference to it with hyperlinks in an information retrieval enviroment.

4.1. A closer look at Prudence

Without any claims of technical thoroughness or complete coverage of all the possible instances, in this section is shown, through a textual fragment and some line of code, the way rules, regular expressions and incremental mark-up cooperate within the software.

```
<BREAK>...attraverso la sentenza della Corte di appello di
Trento - sez. dist. di Bolzano n. 46 del 2008, depositata
il 2 marzo 2008, corretta con ordinanza depositata in
data 11 luglio 2008 e notificata il 9 ottobre 2008.<BREAK>
```

The text above, identified by the sentence splitter module of Prudence, contains two references and represents their context. The authority marker module defines the regular expressions for the detection of authority names and court sections in Italian texts, like:

```
Consiglio = (cons\.|consiglio){S}?{Di}?{S}?(stato|
            ((giust\.|giustizia){S}?(amm\.|
            amministrativa)?)|europeo)

Authority = {Corte}|{Cass}|{Consiglio}|{Tribunale}|
            {Tar}|{Comune}

SecNum = {NumLatino}|{NumOrdinale}

SectionAttr = {Ordinaria}|{Civile}|{Lavoro}|{Penale}|
              {Giuris}|{Tributaria}
SectionDetails = {SectionAttr}|(feriale)|(un(\.|(ite)))|
                 (dist(\.|(accata)))
Section = ({SecNum}?{S}?{Sez})|({Sez}{SS}?
          {SecNum}?)|({SU})|({Sez}{S}?{NumNoDate})
```

Moreover such marker also defines the state 'issuing-aut' and the rules that apply to it. Within the tag <AUT> the authority marker includes even geographical information about the issuing authority, if specified in the text, like the city where a court is located.

```
{Authority}   {output += "<AUT>" + yytext();
               yybegin(ISSUING-AUT);}
<ISSUING-AUT> {
                  {Section}       { output += yytext(); }
                  {SectionAttr}   { output += yytext(); }
                  - cut -
                  {Location}      { output += yytext(); }
                  .               { output += "</AUT>" +
                                    yytext();
                                    yybegin(YYINITIAL); }
              }
```

After date and numbering mark-up, the text fragment has been enriched with several tags:

```
<BREAK>...attraverso la sentenza della <AUT>Corte di appello
di Trento - sez. dist. di Bolzano</AUT> <NUM>n. 46 del
2008</NUM>, depositata il <DATE>2 marzo 2008</DATE>, corretta
con ordinanza depositata in data <DATE>11 luglio 2008</DATE>
e notificata il <DATE>9 ottobre 2008</DATE>.<BREAK>
```

and is given as input to the reference builder, where the regular expressions, which trigger the analysis of a reference to a decision ("sentenza"), ordinance ("ordinanza") and decree ("decreto"), are implemented.

```
Sentenza = (sent(t)?\.|sentenz(a|e))({S}((definitiva)|((di)
             {S}(separazione))))?
OrdDetails = (ingiunzione)|(pagamento)|(rilascio)|(rigetto)|
           (rimessione)|(anticipatori(a|e))|(esecutiv(a|e))|
           (istruttori(a|e))|(ingiuntiv(a|e))|(cautelar(e|i))
Ordinanza = ((ordinanz(a|e))|(ord\.)){S}?(di{S})?{OrdDetails}?
DecretoDetails = (ingiuntiv(i|o))?{S}?(oppost(i|o))?
Decreto = ((decret(o|i))|(d\.i\.)|(d\.{S}?ingiuntivo))
           {S}?{DecretoDetails}?
```

Some rules are defined as well for the'decision' state and shown below as example. The reference builder module tries to fill the metadata of the reference without stepping out of the state 'decision': if this procedure is successful, the reference is considered valid, otherwise, following the parameters defined by the user, the automata can skip the analysis of the current textual instance (*strict configuration*) or extend the research of metadata values outside the 'decision' state (*loose configuration*).

```
{Sentenza}   { reset("sentenza"); saveTipo(yytext());
             source.addText(yytext()); yybegin(DECISION);
             }
<DECISION> {
                {Date}          { saveData(yytext());
                                  source.addText(yytext());}
                {FileNumber}    { saveRuolo(yytext());
                                  source.addText(yytext());}
                {YearNumber}    { saveAnnoNumero(yytext());
                                  source.addText(yytext());}
                {Authority}     { saveEmanante(yytext());
                                  source.addText(yytext());}
                - cut -
                {DecisionMisc}  { source.addText(yytext());}
                .               { source.addText(yytext());
                                  yybegin(YYINITIAL);}
           }
```

The regular expression *DecisionMisc* defines the free text allowed in the 'decision' state (it could also be considered

as *noise* for that particular state). By customizing it upon the corpus of the court of Milan, it has been possible to keep a *strict configuration* parsing.
Finally, the rules of the reference builder trigger the identification of two references: a decision and an order of the court.
In Table 7 the results of the parsing strategies in terms of identified metadata are reported. Moreover the normalization and serialization of such metadata in terms of LEX or ECLI naming conventions are respectively shown.

Metadata	**Values**
Document type:	sentenza [decision]
Issuing authority:	Corte di appello di Trento sez. dist. di Bolzano
Date:	2 marzo 2008
Numero:	n. 46 del 2008
`urn:lex:it:corte.appello.trento:sentenza:2008-03-02;46`	
`ECLI:IT:CAT:2008-03-02:46`	
Document type:	ordinanza [ordinance]
Date:	11 luglio 2008
`urn:lex:it:corte.appello.trento:ordinanza:2008-07-11`	
`ECLI:IT:CAT:2008-07-11`	

Table 7: Parsing results[9]

5. Preliminary tests

A data set of 2487 decisions provided by the Court of Milan has been analysed by legal experts, in order to provide a survey of the most widely used textual citations of cases. To improve the survey coverage, the document dataset has been enlarged by analysing about one thousand more documents gathered from the jurispridential database of the Italian High Court of Cassation[10] as well as of some commercial publishers. This survey collected about 300 different modalities of cases citations, whose lexicon and grammar have been used to create the lexicon and the grammar rules for the parser. The assumption is that such grammar can be reliably used to detect and analyse case references both in the analysed document dataset and outside of it. This assumption is motivated by the wideness of the analysed dataset and by the stability of the number of collected textual reference modalities with the progression of the document dataset analysis.
In the preliminary experiments the parser showed a high reliability for both detecting case citations and extracting the relevant metadata required by the LEX and ECLI naming conventions. In such experiments two main ambiguous type of references have been identified, not currently covered by the parser grammar:

- the case in which the textual components of a citation are separated by one or more subordinate clause, including other references, dates or numbers;
- the case in which a textual component of a citation is implicitly expressed, by making reference to a previous citation (for example, a specific hearing date is indicated by the expression *nella stessa udienza...*[in the same hearing], referring to a previously cited hearing).

[9]assuming that CAT is the ECLI court code for "Corte di Appello di Trento" [Court of Appeal of Trento]
[10]http://www.italgiure.giustizia.it

A possible solution of the first type of ambiguous references can be given by the adoption of a dependency parser, while the second type of them can be faced by analysing the relations between the sentences, rather then performing a sentence-by-sentence text analysis. Considering the low recurrence of such cases, and the complexity and the state-of-the-art dependency parsing accuracy for the Italian language, these solutions have not been implemented yet.

6. Conclusions

Prudence is currently under a systematic test by using a gold standard of case law documents where references are marked-up and represented by the related LEX and ECLI naming conventions. The preliminary tests showed a high reliability of the parsing strategy: in fact, despite of the variability of references wording, the language used in references is highly technical including a limited number of terms and expressions. The reliability of this parsing strategy paves the way for an effective use of this tool to support the daily procedures of document management within courts, as well as support the editorial activities of legal publishers to provide stable and effective services of legal document browsing, retrieval and analysis of specific decisions authoritativeness.

7. References

C. Biagioli, E. Francesconi, P. Spinosa, and M. Taddei. 2005. A legal drafting environment based on formal and semantic xml standards. In *Proceedings of International Conference on Artificial Intelligence and Law*, pages 244–245.

A. Boer, R. Hoekstra, E. de Maat, E. Hupkes, F. Vitali, and M. Palmirani. 2009. Cen workshop agreement 'open xml interchange format for legal and legislative resources'. Technical report, CEN/ISSS Workshop Metalex.

E. de Maat, R. Winkels, and T. van Engers. 2006. Automated detection of reference structures in law. In T. van Engers, editor, *JURIX*, pages 41–50. IOS Press.

EU. 2011. Council conclusions inviting the introduction of the european case law identifier (ecli) and a minimum set of uniform metadata for case law. Technical report, Council of European Union, April. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:127:0001:01:EN:HTML`.

E. Francesconi. 2006. The "norme in rete"- project: Standards and tools for italian legislation. *International Journal of Legal Information*, 34(2):358–376.

E. Francesconi. 2011. Naming legislative resources. In G. Sartor, M. Palmirani, E. Francesconi, and M. Biasiotti, editors, *Legislative XML in the Semantic Web. Principles, Models, Standards for Document Management*, pages 55–80. Springer.

C. Kirchberger. 2011. The 'i' in legal information retrieval. *Vem reglerar informationssamhället? Nordisk årsbok i rättsinformatik*. `http://ssrn.com/abstract=1957558`.

P. Mazzega, D. Bourcier, and R. Boulet. 2009. The network of french legal codes. In *ICAIL*, pages 236–237. ACM.

P. Spinosa, E. Francesconi, and C. Lupo. 2012. A uniform resource name (urn) namespace for sources of law (lex). Technical report, ietf. `http://datatracker.ietf.org/doc/draft-spinosa-urn-lex/`.

M. van Opijnen. 2008. Finding case law on a european scale – current practice and future work. In *Proceedings of the Jurix Conference*, pages 43–52.

R. Winkels, J. de Ruyter, and H. Kroese. 2011. Determining authority of dutch case law. In K.M. Atkinson, editor, *JURIX*, pages 103–112. IOS Press.

Semantic Annotations for Legal Text Processing using GATE Teamware

Adam Wyner, Wim Peters

Department of Computer Science; Department of Computer Science
University of Liverpool; University of Sheffield
Liverpool, UK; Sheffield, UK
adam@wyner.info; W.Peters@dcs.shef.ac.uk

Abstract

Large corpora of legal texts are increasing available in the public domain. To make them amenable for automated text processing, various sorts of annotations must be added. We consider semantic annotations bearing on the content of the texts - legal rules, case factors, and case decision elements. Adding annotations and developing gold standard corpora (to verify rule-based or machine learning algorithms) is costly in terms of time, expertise, and cost. To make the processes efficient, we propose several instances of GATE's Teamware to support annotation tasks for legal rules, case factors, and case decision elements. We engage annotation volunteers (law school students and legal professionals). The reports on the tasks are to be presented at the workshop.

Keywords: Text processing, Legal corpora, Web-based annotation

1. Introduction

Large, public domain corpora of legal texts are increasing available and searchable. Advanced Scholar Search in Google Scholar makes patents, legal opinions, and journals searchable according to: keywords, author, publication, and collections. The searches can be refined by subject areas, court hierarchy, states, and decision date. Each decision is annotated with respect to decisions cited in it, enabling the presentation of a web of citations to be presented[1] WorldLLI, the global, free, independent, and nonprofit organisation of the Legal Information Institutes, offers search in its database of legal texts, generally by keyword and selected database. Similarly, the United Kingdom offers legislation online The National Archives - legislation, where each act contains links to related acts.

There is, plainly, an enormous volume of textual legal material available. To search for complex information or to make use of it in automated processing, the unstructured textual information of sentences, references, and textual presentation must be structured and made machine readable. To do so, we must annotate corpora of texts with semantic annotations (among other potential annotations, e.g. syntactic) and create Gold Standard corpora, which support the development of rule-based or machine learning algorithms that can be used to annotate large volumes of textual data. In some currently available corpora (those mentioned above), such tasks have been carried out, and we find documents with *linked data*, e.g. references within a text are associated with a *URL*, as well as some *metadata*, e.g. data, location, and judicial context. Yet, clearly, there is a great range of information that can be annotated and used.

Recent work by (Maynard and Greenwood, 2012) shows just how far such an approach - semantic annotation of text, creation of a Gold Standard, and development of automated annotation tools - can go. In this study, 42 terabytes of data from the electronic archives of the UK's National Archives were annotated and indexed with respect to a range of elements: dates, government departments and agencies, measurements, a large knowledge base and associated ontology, and so on. Central to the effort was semantic annotation and the creation of a Gold Standard corpus to evaluate the performance of the system; this was created by four domain experts who manually annotated 13 documents from the source corpus using GATE's Teamware to enter and analyse the annotations (Bontcheva et al., 2010).

Related efforts in the legal domain have created annotation tools for smaller corpora evaluated against relatively constrained Gold Standards for arguments (Moens et al., 2007; Mochales-Palau and Moens, 2008; Wyner et al., 2010), elements of legal cases (Francesconi and Pratelli, 2011; Wyner, 2010), rules and norms (de Maat and Winkels, 2010; Wyner and Peters, 2011), and case factors (Ashley and Brüninghaus, 2009; Wyner and Peters, 2010a). Yet, semantic annotation and the creation of Gold Standards is not, in and of itself, straightforward and unproblematic. Generally, a small number of annotators are deployed on a fragment of the corpus due to the cost and complexity of the task. Moreover, annotation guidance and adjudication are significant issues (Maeda et al., 2008).

In view of the problems and limitations of current annotation campaigns, we suggest a means to broaden participation of annotators. This will allow us to annotate more text with more semantic annotations, leading to higher quality, richer Gold Standard corpora. To do this, we use GATE's Teamware to support annotation tasks for legal rules, case factors, and case decision elements. We engage annotation volunteers (law school students and legal professionals), who are domain specialists. We exercise the tool on a corpus of texts appropriate to the domain - regulations and intellectual property decisions. In the following, we outline GATE's Teamware. In section 3., we mention the approach to annotators, guidelines, and evaulation. The annotations

[1] Accessed April 2, 2012. Search for exact phrase *intellectual property* among legal opinions in California Advanced Scholar Search returns, among others, *Moore v Regents of University of California*. The corpus is based on law.resource.org, which offers bulk access to primary legal materials.

and corpora we work with are discussed in section 4. . We close with a sample screenshot from a previous online annotation exercise. We report on the results of the current compaign at the SPLeT 2012 workshop.

2. Teamware

To create high quality, annotated corpora, we need a clear methodology, guidelines for annotators, a means to serve text and annotation tools to annotators, storage of the annotated texts, measures for inter-annotator agreement, and adjudication of annotations (Maynard and Greenwood, 2012). Teamware provides a unified enviroment to carry out these various tasks. The tool is web-based, so no local installation of software is required, and the data is stored in a central repository. The tool supports a range of roles (e.g., annotators, editors, managers) appropriate to different actors and phases of the annotation process, allowing non-specialists to participate in the annotation task. On the other hand, expert curators can then adjudicate the gathered annotations. In addition to supporting users in annotating text, Teamware uses GATE components to *preannotate* the text for a range of annotations, which relieves the annotator of some aspects of the annotation task. Business process statistics are kept on the tasks, representing time each annotator spends per document, percentage of completed documents, and other measures.

3. Annotators, Guidelines, and Evaluation

For annotators, we propose to work with contacts in law schools and legal societies to engage them voluntarily in a collaborative task that is similar to the annotating task they already individually engage in to *brief cases*, but using an online tool to annotate, compare, and evaluate corpora created by the users. In future work, we look forward to tying together more closely the annotation tasks with learning objectives in law schools, for example, by using the tool to support legal case analysis and comparison as a basis for student discussion.

To support the annotators in their task, they must be receive guidance. In a small pilot study for online annotation of legal case factors, we provided instructions on how to access and use the tool itself as well as information on the annotations to be identified in the text, e.g. Legal Case Annotation. We expect to extend and expand these instructions for more widespread use.

The tool supports multiple annotators who are annotating the same text with the same annotation set. Thus, a text is multiply annotated and can be compared for *interannotator agreement*, the extent to which annotators agree not just on the selection of annotations on the text, but the exact textual span covered by the annotation. GATE Teamware provides tools for measuring interannotator agreement. In addition, there is an *adjudication* tool so that differences between annotators can be decided in favour of the *correct* or *consistent* annotation. In this way, GATE Teamware supports the development of *Gold Standards*.

4. Annotations and Corpora

The target annotations are based on prior work for each of the following topic areas. As each of the sub-topics may have a large range of possible annotations, we make a selection of relevant annotations as a basis for further systematic and controlled development of use of the tool.

4.1. Legal Case Factors

To facilitate legal case-based reasoning, the legal case factors must be analysed. We focus on cases concerning *intellectual property* and factors discussed in the cases. For a corpus, we have 140 cases that have been used in the CATO analysis of legal cases (Aleven, 1997). The factors are expressed in (Wyner and Peters, 2010b), which are then further decomposed in (Wyner and Peters, 2010a). From (Aleven, 1997), we have 27 base level factors such as follows, where we have an index F1, a label *Plaintiff-disclosed-information-in-negotiations*, the side of the dispute that the factor favours, a description of the factor, and comments on when the factor does and does not apply. The latter three elements can be used to aid the annotator.

- F1 Plaintiff-disclosed-information-in-negotiations
- Favours defendant.
- Plaintiff disclosed information during negotiations with defendant. The defendant fairly obtained the information and the plaintiff was not interested to maintain the information as a secret.
- Applies if the plaintiff disclosed the information to defendant during negotiations for a joint venture, licensing agreement, sale of a business, etc..
- Does not apply if the defendant learned the information while employed by plaintiff.

Among others, we have the following factors:

- F6 Plaintiff-adopted-security-measures
- F7 Defendant-hired-plaintiff-employee
- F10 Plaintiff-disclosed-information-to-outsiders
- F21 Defendant-knew-information-confidential
- F27 Plaintiff-disclosed-information-in-public-forum

In this task, the objective is for the annotator to annotate the sentence or sentences that indicate the relevant factor in the case decision.

4.2. Rules

For the analysis of regulations and legislation, it would be very useful to identify, extract, and process *legal rules*. This part of the task is based on (Wyner and Peters, 2011). As an initial basis, we use a corpus of passages from US Code of Federal Regulations for blood banks on testing requirements for communicable disease agents in human blood. This is a four page document of 1,777 words. The model of analysis proposed includes annotation for:

- Agent and theme, which are semantic roles that must be associated with noun phrases in grammatical (subject or object) roles in the sentence. These are used to account for active-passive alternations and identify the individual's relationship to the deontic concept.
- Deontic modals and verbs.
- Main verbs.
- Exception clauses, which may appear in lists.
- Conditional sentences along with their antecedents and consequences. Antecedents may appear in lists.

4.3. Case Elements

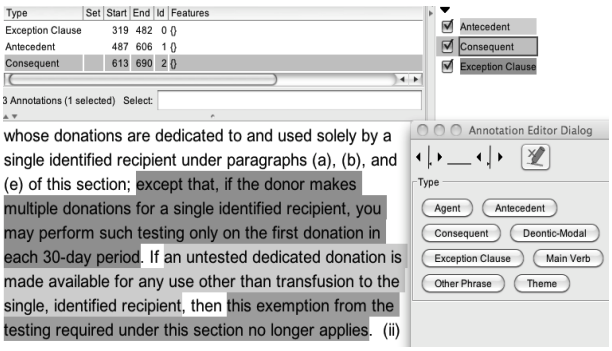
In addition to factors relevant to legal case-based reasoning, we are interested to identify and extract for further processing a range of elements that appear in legal cases (Wyner, 2010; Francesconi and Pratelli, 2011). We use some of the cases for the CATO case base. Among the elements of interest are:

- Case citation, cases cited, precedential relationships.
- Names of parties, judges, attorneys, court sort....
- Roles of parties, meaning plaintiff or defendant, and attorneys, meaning the side they represent.
- Final decision.
- Case structural features such as sections.
- Causes of action.

4.4. A Sample

In Figure 1, we have a screen shot of Teamware after a user has annotated parts of the document she has been served. The annotator receives a document in the online tool, highlights a passage, selects an annotation from the *Annotation Editor Dialog*, then moves on to annotate the next passage. In the sample, we have annotations (in colour in the original, but in greyscale in this paper) for an exception clause *except that...30-day period*, an antecedent of a conditional *an untested...recipient*, and the consequent of a conditional *this exemption...applies.* Once annotated by several annotators, we can evaluate interannotator agreement, export the annotated information in XML, and further process it.

Figure 1: Sample of Teamware to Annotate Rules



5. Conclusion

In this paper, we have briefly outlined motivation, background, tool, corpora, and target annotations that we study in the annotation exercise. The results of the exercise are to be reported at the SPLeT 2012 meeting, which is part of LREC 2012. We expect that this will be the start of a broader movement to *crowdsource* legal text analytics and semantic analysis on a larger scale, which will yield greater understanding of and use for legal information.

6. Acknowledgements

The first author was supported by the FP7-ICT-2009-4 Programme, IMPACT Project, Grant Agreement Number 247228. The views expressed are the author and should not be taken as representative of the project.

7. References

Vincent Aleven. 1997. *Teaching case-based argumentation through a model and examples*. Ph.D. thesis, University of Pittsburgh.

Kevin D. Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intellence and Law*, 17(2):125–165.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, and Valentin Tablan. 2010. Web-based collaborative corpus annotation: Requirements and a framework implementation. In *Proceedings of New Challenges for NLP Frameworks*, Malta, May.

Emile de Maat and Radboud Winkels. 2010. Automated classification of norms in sources of law. In Enrico Francesconi, et al., editors, *Proceedings of SPLeT '10*, pages 170–191. Springer.

Enrico Francesconi and Tommaso Pratelli. 2011. A twofold parsing strategy for italian court decisions. In Katie Atkinson, editor, *Proceedings of JURIX '11*, pages 125–129. IOS Press.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. 2008. Annotation tool development for large-scale corpus creation projects at the linguistic data consortium. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC '08*. European Language Resources Association.

Diana Maynard and Mark Greenwood. 2012. Large scale semantic annotation, indexing and search at the national archives. In *Proceedings of LREC '12*. European Language Resources Association. To appear.

Raquel Mochales-Palau and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In Enrico Francesconi, et al., editors, *Proceedings of JURIX '08*, pages 11–20. IOS Press.

Marie-Francine Moens, Erik Boiy, Raquel Mochales-Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of ICAIL '07*, pages 225–230. ACM Press.

Adam Wyner and Wim Peters. 2010a. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In Radboud Winkels, editor, *Proceedings of JURIX '10*, pages 127–136. IOS Press.

Adam Wyner and Wim Peters. 2010b. Towards annotating and extracting textual legal case factors. In *Proceedings of SPLeT '10*, Malta. To appear.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *Proceedings of JURIX '11*, pages 113–122. IOS Press.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Proceedings of SPLeT '09*, pages 60–79. Springer.

Adam Wyner. 2010. Towards annotating and extracting textual legal case elements. *Informatica e Diritto: Special Issue on Legal Ontologies and Artificial Intelligent Techniques*, 19(1-2):9–18.

Legal Information Extraction ← Machine Learning Algorithms + Linguistic Information

Paulo Quaresma

Computer Science Department & CENTRIA – AI Centre
School of Sciences and Technology,
University of Évora, Portugal
pq@di.uevora.pt

Abstract

In order to automatically extract information from legal texts we propose the use of a mixed approach, using linguistic information and machine learning techniques. In the proposed architecture, lexical, syntactical, and semantical information is used as input for specialized machine learning algorithms, such as, support vector machines. This approach was applied to collections of legal documents and the preliminary results were quite promising.

Keywords: Information Extraction, Semantic Analysis, Machine Learning Algorithms

1. Introduction

Information extraction from text documents is an important and open problem. Although this is a general domain problem, it has a special relevance in the legal domain. For instance, it is very important to be able to automatically extract information from documents describing legal cases and to be able to answer queries and to find similar cases.
Much research work on this topic has been done in the last years, as it is described, for instance, in Stranieri and Zeleznikow's book "Knowledge Discovery from Legal Databases" (Stranieri and Zeleznikow, 2005). Typical approaches vary from machine learning techniques, applied to the text mining task, to the use of natural language processing tools.

2. Proposal

We claim that a mixed approach, using deep linguistic information and machine learning techniques, is the best approach to handle this problem and to obtain good results.
By "deep linguistic information" we mean lexical, syntactical and semantical information, linked with an ontology representing the knowledge of the domain.
The overall idea is to use natural language processing tools to analyse the legal texts and to obtain:

- Lexicon with part-of-speech (POS) tags;
- Syntactical parse trees;
- Partial semantical representation.

This linguistic information can be used as input features for specialized machine learning algorithms, which will be responsible for high-level information tagging and extraction. As machine learning techniques we propose the use of kernel-based ones, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), which are able to handle complex structured data as input.
The extracted information – mainly legal concepts and named entities – can be used to populate domain ontologies, allowing the enrichment of documents and the creation of high-level legal information retrieval systems. These legal information systems are "semantic-aware" ones and they are able to answer queries about concepts, entities and events.

3. Example

As an example, suppose the legal text has the following (simple) sentence:

- The judge decided in favor of the plaintiff.

As natural language processing tools we have used NLTK – Natural Language Toolkit – from the University of Pennsylvania, "C&C" for the syntactical parser and "Boxer" for the syntactic to semantic representation (Curran et al., 2007; Bos, 2008).
Applying these tools to the presented example[1], we'll obtain the following parse tree:

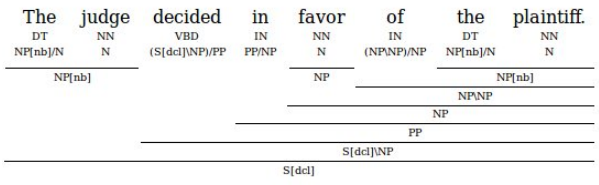


Figure 1: Parse tree

In this tree it is possible to identify the noun phrase *The judge*, the main verb *decided* and the prepositional phrase *in favor of the plaintiff*.
Applying the "Boxer" tool to the output of the C&C parser, we'll obtain the discourse structure represented in figure 2.
In this structure we have:

- two entities *x0* and *x1*, which represent the *judge* and the *plaintiff*;
- the event *x2*, which represents the main action *to decide* and has an agent *x0*, the judge, and an object *x3*;

[1] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Demo

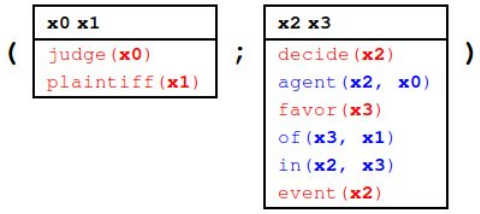


Figure 2: Discourse representation structure

- an object of the decision – *x3* –, which is related with *x2* and *x1* through the relations *in* and *of*.

As it was shown, the obtained output allows the automatic identification of entities and actions and the inference of their relations.
This information can be added as input features to machine learning algorithms aiming to improve the results of the named entity recognition task (NER) – identification of persons, organizations, and places.
Moreover, being a logically-based formalism, this representation allows an "easier" implementation of knowledge-based systems. In the presented example, suppose we have a legal ontology with the concept/class *person* with subclasses *judge* and *plaintiff*; it is possible to automatically populate these subclasses with instances *x0* and *x1*. The same approach can be applied to automatically populate instances of *events* and to create links between the created instances (as described in (Saias and Quaresma, 2005)). The obtained ontology (classes, instances, and relations) can be represented in a web language, such as OWL – Ontology Web Language, allowing its access through specialized semantic web search engines or using a query language, such as SPARQL (SPARQL Protocol And RDF Query Language).

4. Experiences

The use of deep linguistic information to automatically create and to populate legal ontologies was proposed and described in (Saias and Quaresma, 2005). This approach was completely based on symbolic natural language processing tools and it was applied to a collection of documents from the Portuguese Attorney General's Office. The limitations of the existent NLP tools (parsers, semantic analyzers) were one of the major reasons we've extended this approach to also use machine learning techniques.
In another previous work (GonÃ§alves and Quaresma, 2010) we have partially applied this methodology to a corpus of legal documents from the EUR-Lex site[2] within the "International Agreements" sections and belonging to the "External Relations" subject. The obtained results were very promising and, for the main concepts identification task, we obtained values higher than 95% for the precision and 90% for the recall. The identification of named entities (NER) showed also good results, varying from error rates of 0.1% for dates to around 15% in the identification of places and to a high value of 65% for organizations (due to specific problems reported in the cited paper). However, this work represented a first approach and our proposal was not fully applied, as we didn't use semantic information.
In another more recent work (Gaspar et al., 2011) we've applied the complete "deep" linguistic analysis to a collection of texts from the Reuters dataset, obtaining a partial semantic representation of the sentences – as discourse representation structures (DRSs). These structures were represented by direct graphs. We have also proposed a graph kernel function able to calculate the similarity of the semantic structures and we used Support Vector Machines to classify texts. The results were promising with an accuracy higher than 50%. As referred in the previous section, as natural language processing tools we used the "C&C" syntactical parser and "Boxer" to obtain the semantic representation (Curran et al., 2007; Bos, 2008).

[2] http://eur-lex.europa.eu/en/index.htm

5. Conclusions

We proposed to use deep linguistic information and machine learning techniques to the legal information extraction task. The results obtained in preliminary results were quite promising. However, it is necessary to perform more experiences with bigger legal text collections.

6. References

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

Miguel Gaspar, Teresa Gonçalves, and Paulo Quaresma. 2011. Text classification using semantic information and graph kernels. In *EPIA-11, 15th Portuguese Conference on Artificial Intelligence*, pages 790–802, Lisbon, PT, October. ISBN: 978-989-95618-4-7.

Teresa Gonçalves and Paulo Quaresma. 2010. Using linguistic information and machine learning techniques to identify entities from juridical documents. In E. Francesconi, E. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts*, Lecture Notes in Artificial Intelligence 6036, pages 44–59. Springer.

Jose Saias and Paulo Quaresma. 2005. A methodology to create legal ontologies in a logic programming information retrieval system. In R. Benjamins, P. Casanovas, A. Gangemi, and B. Selic, editors, *Law and the Semantic Web*, Lecture Notes in Computer Science LNCS 3369, pages 185–200. Springer-Verlag.

A. Stranieri and J. Zeleznikow. 2005. *Knowledge Discovery from Legal Databases*. Law and Philisohy Library. Springer.

Problems and Prospects in the Automatic Semantic Analysis of Legal Texts
A Position Paper

Adam Wyner

Department of Computer Science, University of Liverpool
Ashton Building, Asthon Street, Liverpool, L69 3BX, United Kingdom
adam@wyner.info

Abstract

Legislation and regulations are expressed in natural language. Machine-readable forms of the texts may be represented as linked documents, semantically tagged text, or translation to a logic. The paper considers the latter form, which is key to testing consistency of laws, drawing inferences, and providing explanations relative to input. To translate laws to a machine-readable logic, sentences must be parsed and semantically translated. Manual translation is time and labour intensive, usually involving narrowly scoping the rules. While automated translation systems have made significant progress, problems remain. The paper outlines systems to automatically translate legislative clauses to a semantic representation, highlighting key problems and proposing some tasks to address them.

Keywords: Translation, Semantics, Syntax, Text Analysis

1. Introduction

Laws as found in legislation and regulations are expressed in natural language. To make laws automatically processable, they must be made machine-readable since the language of the law is, from the point of view of a computer, an unstructured sequence of characters. There are several approaches to making legal texts machine-readable, depending on the goals and purposes to be served by the processed text. Among the approaches, legal texts may be processed to link documents, to annotate for information extraction, and to parse and translate them to a logic. Each of the approaches has its own use. For linked documents, the objective is to identify components in the text that may be associated with some other web-based document. For example, references to a law in a text may be associated with a web-accessible link to the particular law or other further information, e.g. The British Nationality Act 1981. Google's Advanced Scholar Search facility allows searches restricted to terms in legal opinions and returns decisions that have links to cases cited in the decision, e.g. Advanced Micro Devices, Inc. v. Intel Corp.. Linking documents not only helps to identify related documents, but also to highlight *relationships* between texts; in legal decisions, case citations can be used to indicate the relevance of precedents. In another approach, legal documents can be automatically tagged with a variety of sorts of annotations to enable information extraction and fine-grained search in the body of legal documents (Maynard and Greenwood, 2012; Wyner and Peters, 2011). In the final approach, legal texts are processed and rendered into a machine-processable logic that can be used for testing consistency of laws, drawing inferences, and giving users meaningful explanations following a consultation. While the first two approaches have seen very rapid, widespread, and continuing development, the third has not, despite being one of the early achievements in AI and Law, its current commercial success, and well-developed NLP tools.

This position paper is a pointer to problems and prospects bearing on automatic translation of legal text from natural language to a machine-processable logic. We begin with some background, then turn to some aspects of state-of-the-art systems on semi-automated systems, and finally consider a fully automated system. The discussion is illustrated with a well-known working example.

2. Background - Manual Translation

One of the early ambitions and achievements of artificial intelligence and law was to formalise legislation as a logic program. Several large scale projects were carried out (Sergot et al., 1986; Bench-Capon et al., 1987; Sergot, 1988). The method, carried out manually, was to take the source legal text, identify the relevant textual portions, decompose and paraphrase them as necessary, and then formalise the language in an executable logic such as Prolog, creating an *expert system*. From this formalisation, ground facts may be provided to the system which are then used to draw inferences and the rule system could be tested for consistency. Translation of the *British Nationality Act 1981* was one such exercise. The first clause is as follows. It is stated in the act that "after commencement" means on or after the date when the act comes into force.

> l.-(l) A person born in the United Kingdom after commencement shall be a British citizen if at the time of birth his father or mother is (a) a British citizen; or (b) settled in the United Kingdom.

The clause is translated into Prolog as:

```
is_a_British_citizen(X) :-
    was_born_in_the_U.K.(X), was_born_on_date(X,Y),
    is_after_or_on_commencement(Y),
    has_a_parent_who_qualifies_under_1.1_on_date(X,Y).
```

This is a "first" draft translation, and the literature discusses a range of issues that must be addressed such as dependencies between subsequent portions of text, the introduction of negation, drawing out implicit information, the complex structure of the clauses. An overall point is that the axioms of the legislation are formulated from the source by a

methodology of trial and error; that is, there is no systematic or automated analysis of the natural language text. Not only does this make the analysis expensive to produce and maintain, but does not facilitate reuse as each predicate is *sui generis* rather than composed from linguistic modules. Nonetheless, the translation provides a *gold standard*: in an interactive environment, a user queries the system, answers questions, and receives determinations and explanations.

3. Manually Paraphrased and Automatically Translated

Since this early work, some commerical products have become available which support aspects of this process and serve the resultant expert systems to users on the web (Johnson and Mead, 1991), (Dayal et al., 1993), (Dayal and Johnson, 2000). In particular, Oracle Policy Management can take rules from legislation in natural language and automatically translate them into a logic; an inference engine is applied to grounded statements, providing determinations; there is a web-interface to serve the system statements to users. Explanatory notes, document access, and alternative evaluations are auxiliary capabilities. It has been applied to the examples discussed in (Sergot et al., 1986) and many other acts; it is in widespread use by government agencies in the United Kingdom and United States, e.g. for tax calculation and citizen benefits.

While this is a very significant development, its overall contribution is limited in two respects. First, the methodology of analysis, though industrialised, remains largely that of trial and error: the source text is analysed manually, scoped, and paraphrased *in a controlled natural language (fixed grammatical constructions) to meet the constraints of the parser and semantic interpreter*; the system uses "just enough" natural language processing to satisfy the clients' requirements. Second, as a proprietary product, the system is restricted in exposure, use, and development.

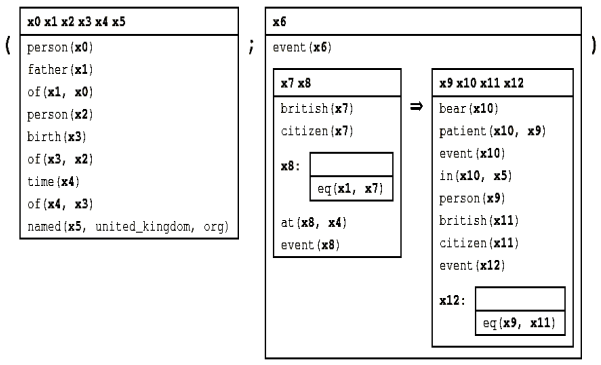
As an alternative, *Attempto Controlled English* (ACE) is a controlled language (Fuchs et al., 2008; Wyner et al., 2009) that has been applied in some small, non-legal domains (Shiffman et al., 2010; Wyner et al., 2010). As a controlled language, any source text must be paraphrased to fulfil the specifications of the language. Given the sentence, the system automatically parses and semantically represents an unambiguous semantic formula in *Segmented Discourse Representation Theory* (SDRT), which can be input to an inference engine. As discussed in (Wyner et al., 2010), matters are not straightforward for one must carefully evaluate whether the output semantic formula accurately represents the intended meaning of the input sentence, adjusting the input sentence accordingly. It also remains to be seen whether the parser and semantic interpreter can process the sorts of sentences we find in legal texts. ACE is not yet associated with an expert system interface. Despite these issues, the system has several advantages: the input is in natural language, the result is a single, unambiguous semantic translation, the system is open source and extensible, it has a web interface, and it is already highly flexible.

4. A Fully Automatic System

A more expressive, flexible, and powerful natural language processing system is C&C/Boxer (Bos et al., 2004); it has extensive, efficient parsing using categorial grammar and a translation to SDRT; while it may be controlled, it is not constrained to be so. It generates the most *likely* parse and semantic representation, requiring analysis of the results, selection among the alternative analyses, or modification of the input till one gets the intended representation.

We illustrate the output of the tool with a paraphrase of our example legislative clause to show the results and the issues. For example, *A person born in the United Kingdom shall be a British Citizen if at the time of his birth his father is a British citizen.* has the output representation in Figure 1. We found we had to make explicit the implicit pronominal relationships and also the gender. The semantics does not specify the relation between the man and time of birth or between the man and the father. And we note that we have only represented a small fragment.

Figure 1: SDRT of Sample Statement



Even from this small sample, we can see that the problems of manual translation have shifted from initial analysis to evaluation of output - the discourse elements and predicates. The first problem is that x0 (the person with the father) and x2 (the person with the birthday) are possibly distinct. In the antecedent of the conditional, x1 (the father) is a British Citizen at the time of the birth of x2. But, x2 need not be identical to the person who is the son of that citizen. Finally, in the conclusion, x9 gets British Citizenship, but is not identified with either x0 or x2.

Despite these problems, automated systems are systematic, grounded, and open to refinement. Besides continued evaluation of output, automated systems need testing suites to support evaluation, especially where large, complex expressions and documents are concerned. Furthermore, the parser and semantic interpreter must developed and refined to meet the requirements found in legal textual language.

5. Conclusion

In this position paper, we have briefly presented three main approaches to semantic representation of legislative documents that can be used for automated inference. Some of their problems and prospects have been outlined with the intention that the observations can be used to develop more sophisticated systems.

6. Acknowledgements

The author was supported by the FP7-ICT-2009-4 Programme, IMPACT Project, Grant Agreement Number 247228. The views expressed are the author.

7. References

Trevor Bench-Capon, George Robinson, Tom Routen, and Marek Sergot. 1987. Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation. In *Proceedings of ICAIL '87*, pages 190–198. ACM.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *Proceedings COLING '04*, pages 1240–1246, Morristown, NJ, USA. Association for Computational Linguistics.

Surend Dayal and Peter Johnson. 2000. A web-based revolution in australian public administration. *Journal of Information, Law, and Technology*, 1. Online.

Surendra Dayal, Michael Harmer, Peter Johnson, and David Mead. 1993. Beyond knowledge representation: commercial uses for legal knowledge bases. In *Proceedings of ICAIL '93*, pages 167–174. ACM.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2008. Attempto controlled english for knowledge representation. In Cristina Baroglio, et al., editors, *Reasoning Web*, pages 104–124. Springer.

Peter Johnson and David Mead. 1991. Legislative knowledge base systems for public administration: Some practical issues. In *Proceedings of ICAIL '91*, pages 108–117. ACM.

Diana Maynard and Mark Greenwood. 2012. Large scale semantic annotation, indexing and search at the national archives. In *Proceedings of LREC '12*, Istanbul, Turkey, May.

Marek Sergot, Fariba Sadri, Robert Kowalski, Frank Kriwaczek, Peter Hammond, and Therese Cory. 1986. The British Nationality Act as a logic program. *Communications of the ACM*, 29(5):370–386.

Marek Sergot. 1988. Representing legislation as logic programs. *Machine Intelligence*, pages 209–260.

Richard N. Shiffman, George Michel, Michael Krauthammer, Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2010. Writing clinical practice guidelines in controlled natural language. In Norbert E. Fuchs, editor, *Proceedings of the 2009 conference on Controlled Natural Language*, pages 265–280. Springer.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *Proceedings of JURIX '11*, pages 113–122. IOS Press.

Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert E. Fuchs, Stefan Höfler, Ken Jones, Kaarel Kaljurand, and Tobias Kuhn. 2010. On controlled natural languages: properties and prospects. In Norbert E. Fuchs, editor, *Proceedings of the 2009 conference on Controlled Natural Language*, volume 5972 of *Lecture Notes in Computer Science*, pages 281–289. Springer.

Adam Wyner, Tom van Engers, and Kiavash Bahreini. 2010. From policy-making statements to first-order logic. In Kim Normann Andersen, Enrico Francesconi, Åke Grönlund, and Tom M. van Engers, editors, *Proceedings of EGOVIS '10*, pages 47–61. Springer.

The SPLeT–2012 Shared Task on Dependency Parsing of Legal Texts

Felice Dell'Orletta*, Simone Marchi*, Simonetta Montemagni*, Barbara Plank†, Giulia Venturi◇

* Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
† DISI, University of Trento, Italy
◇ Scuola Superiore Sant'Anna di Studi Universitari e di Perfezionamento, Pisa, Italy
{felice.dellorletta,simone.marchi,simonetta.montemagni,giulia.venturi}@ilc.cnr.it, barbara.plank@disi.unitn.it

Abstract

The 4th Workshop on "Semantic Processing of Legal Texts" (SPLeT–2012) presents the first multilingual shared task on Dependency Parsing of Legal Texts. In this paper, we define the general task and its internal organization into sub–tasks, describe the datasets and the domain–specific linguistic peculiarities characterizing them. We finally report the results achieved by the participating systems, describe the underlying approaches and provide a first analysis of the final test results.

Keywords: Domain Adaptation, Dependency Parsing, Legal Text Processing

1. Introduction and Motivation

As overtly claimed by McCarty (2007), "one of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier". This entails that it is of paramount importance to use Natural Language Processing (NLP) techniques and tools that automate and facilitate the process of knowledge extraction from legal texts. In particular, it appears that a number of different legal text processing tasks could benefit significantly from the existence of dependency parsers reliably dealing with legal domain texts, e.g. automated legal reasoning and argumentation, semantic and cross–language legal information retrieval, document classification, legal drafting, legal knowledge discovery and extraction, as well as the construction of legal ontologies and their application to the legal domain.

Dependency parsing thus represents a prerequisite for any advanced IE application. However, since Gildea (2001) it is a widely acknowledged fact that state–of–the–art dependency parsers suffer from a dramatic drop of accuracy when tested on domains outside of the data from which they were trained or developed on. In order to overcome this problem, the last few years have seen a growing interest in developing methods and techniques aiming at adapting current parsing systems to new domains. This is testified by several initiatives organized around this topic: see, for instance, the "Domain Adaptation Track" organized in the framework of the CoNLL 2007 Shared Task (Sagae and Tsujii, 2007a), or the ACL Workshop on "Domain Adaptation for Natural Language Processing" (DANLP, 2010). In this context, a particularly relevant initiative is represented by the "Domain Adaptation Track" (Dell'Orletta et al., 2012) organized in the framework of the third evaluation campaign of Natural Language Processing and Speech tools for Italian, Evalita–2011[1], where participants were asked to adapt their dependency parsing systems to the legal domain.

With the only exception of the Evalita–2011 "Domain Adaptation Track" whose results provided relevant feedback in this direction (unfortunately circumscribed to the Italian language), so far very few attempts have been carried out to quantify the performance of dependency parsers on legal texts (e.g. law or case law texts). Among the reasons behind this lack of attention is the unavailability of gold corpora of legal texts annotated with syntactic information with respect to which such an evaluation could be carried out. To our knowledge, exceptions exist only for German and Italian (as mentioned above). The first is the case of the corpus including 100 sentences taken from German court decisions and syntactically manually annotated, as described by Walter (2009). However, this corpus is currently encoded following the PReDS parser (Braun, 2003) native annotation format; its exploitation for the evaluation of dependency parsers would require the conversion of the native PReDS annotation into some kind of standard representation format (e.g. CoNLL).

For the Italian language two different annotated corpora exist: *i)* the portion of the Turin University Treebank (TUT)[2], developed at the University of Torino, including a section of the Italian Civil Law Code (28,048 tokens; 1,100 sentences) annotated with syntactic dependency information and *ii)* TEMIS (Venturi, 2012), a corpus of legislative texts (15,804 tokens; 504 sentences) enacted by three different releasing agencies (i.e. European Commission, Italian State and Piedmont Region) and regulating a variety of domains which is annotated with syntactic and semantic information. Interestingly, the two corpora represent two different sub–varieties of the Italian legal language. According to one of the main Italian scholars of legal language Garavelli (2001), the Civil Law Code articles are less representative of the much cited linguistic complexity of the so–called Italian *legalese* (i.e. the variety of Italian used in the legal domain) with respect to other kinds of legislative texts such as laws, decrees, regulations, etc. This is confirmed by the results achieved in the "Dependency Parsing" Track of Evalita–2011 (Bosco and Mazzei, 2012) where all participant parsers have shown better performances when tested on the Italian Civil Law Code test set than when tested on the newspapers test corpus. Further evidence in the same

[1] http://evalita.fbk.eu/index.html

[2] http://www.di.unito.it/∼tutreeb/

direction emerged within the the Evalita–2011 "Domain Adaptation Track" (Dell'Orletta et al., 2012), where a subset of TEMIS was used: it turned out that parsing systems need to be adapted to reliably analyse legal texts such as laws, decrees, regulations, etc.
Following these premises, the shared task organised in the framework of the 4th Workshop on "Semantic Processing of Legal Texts" (SPLeT–2012) on dependency parsing of legal texts was aimed at: providing common and consistent task definitions and evaluation criteria in order to identify the specific challenges posed by the analysis of this type of texts across different languages; obtaining a clearer idea of the current performance of state–of–the–art parsing systems; and last but not least, developing and sharing multilingual domain–specific resources.

2. Definition of the Task

The shared task was organised into two different subtasks as described below:

1. **Dependency Parsing**: this represents the basic and mandatory subtask, focusing on dependency parsing of legal texts, aimed at testing the performance of general parsing systems on legal texts;

2. **Domain Adaptation**: this is a more challenging (and optional) subtask, focusing on the adaptation of general purpose dependency parsers to the legal domain, aimed at investigating methods and techniques for automatically extracting knowledge from large unlabelled target domain corpora to improve the performance of general parsing systems on legal texts.

The languages dealt with are English and Italian. Evaluation has been carried out in terms of standard accuracy dependency parsing measures, i.e. labeled attachment score (LAS) including punctuation, with respect to a test set of texts from the legal domain.

3. Datasets

For both languages, different datasets have been distributed. For the source domain, task participants have been provided with *i)* a training set exemplifying general language usage and consisting of articles from newspapers and *ii)* a manually annotated development set, also including labeled dependency relations. For the target domain, they have been supplied with *i)* a target corpus including automatically generated sentence splitting, tokenization, morpho–syntactic tagging and lemmatization, and *ii)* a development set, as for the source domain.
All distributed data adhere to the CoNLL 2007 tabular format used in the Shared Task on Dependency Parsing (Nivre et al., 2007) and they are described in detail in the following two sections.
Note that whereas for both English and Italian the final test set is represented by legislative texts enacted by the European Commission (namely, the English and Italian version of the same texts), the domain of development corpora is different for the two languages: for English the development corpora are represented by biomedical abstracts, for Italian they include legal texts belonging to a different sub–variety of the legal language. This is in line with the experimental setup defined for the "Domain Adaptation Track" of the CoNLL 2007 Shared task, where participants were provided with biomedical abstracts as development data, and chemical abstracts and parent–child dialogues as two separate sets of test data.

3.1. Italian Dataset

For the Italian language, the source domain data is drawn from a corpus of news, i.e. the ISST–TANL corpus jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) and the University of Pisa, exemplifying general language usage and consisting of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). This corpus has already been used in the Evalita–2011 "Domain Adaptation Track" (Dell'Orletta et al., 2012). Two different datasets have been distributed to participants: a training corpus (hereafter referred to as *it_isst_train*) of 71,568 tokens and 3,275 sentences and a test corpus (hereafter referred to as *it_isst_test*) of 5,175 tokens (231 sentences).
As target domain data, two different sets have been distributed:

1. a set used as development data drawn from an Italian legislative corpus, gathering laws enacted by Italian State and Regions and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:

   (a) a corpus of 13,095,574 tokens and 660,293 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;

   (b) a manually annotated test set, also including labeled dependency relations, consisting of 5,194 tokens and 118 sentences (hereafter referred to as *it_NatRegLaw*);

2. a set used as test data drawn from an Italian legislative corpus, gathering laws enacted by European Commission and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:

   (a) a corpus of 28,263,250 tokens and 1,300,451 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;

   (b) a manually annotated test set, i.e. sentence–splitted, tokenized, morpho–syntactically tagged and lemmatized, consisting of 5,662 tokens and 241 sentences (hereafter referred to as *it_gold_EULaw*).

The source and target domain data are annotated according

to the morpho–syntactic[3] and dependency[4] tagsets jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project[5].

3.1.1. Source vs Target Domain Corpora Annotation Criteria

Note that in order to properly handle legal language peculiarities, annotation criteria have been extended to cover domain–specific constructions. The specializations are concerned with both sentence splitting and dependency annotation.

For sentence splitting, in the target domain corpora sentence splitting is overtly meant to preserve the original structure of the law text. This entails that also punctuation marks such as ';' and ':', when followed by a carriage return, are treated as sentence boundary markers.

For what concerns dependency annotation, it should be considered that legal texts are characterized by syntactic constructions hardly or even never occurring in the source domain corpora. In order to successfully cope with such peculiarities of legal texts, dependency annotation criteria have been extended to cover the annotation of *a)* elliptical constructions, *b)* participial phrases as well as *c)* long distance dependencies resulting in non–projective links, to mention only a few. All these peculiar constructions have been explicitly represented in the development and final test sets.

3.2. English Dataset

For the English language, the source domain data is represented by the training and test data distributed in the CoNLL 2007 Shared Task. The two sets of data were extracted from the Penn Treebank (PTB)[6] which consists, according the description provided by the Linguistic Data Consortium[7], of 2,499 stories selected from a three year Wall Street Journal (WSJ) collection of 98,732 stories for syntactic annotation. In more detail, the distributed training set (hereafter referred to as *english_ptb_train*) includes sections 02–11 of the WSJ and is a corpus of 446,573 tokens and 18,577 sentences; the test set (hereafter referred to as *english_ptb_test*) is a subset of section 23 of the WSJ for a total amount of 5,003 tokens and 214 sentences.

As target domain data, two different sets have been distributed:

1. a development data set, including the files used for the final testing of the systems in the "Domain Adaptation Track" of the CoNLL 2007 Shared task, namely:
   (a) a corpus of chemical abstracts (CHEM corpus) of 10,482,247 tokens and 396,128 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;
   (b) a manually annotated test set, also including labeled dependency relations, consisting of 5,001 tokens and 195 sentences (hereafter referred to as *english_pchemtb*);
2. a test data set, drawn from an English legislative corpus gathering laws enacted by the European Commission and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:
   (a) a corpus of 25,942,241 tokens and 1,260,621 sentences automatically splitted, tokenized, morpho–syntactically tagged and lemmatized;
   (b) a manually annotated test set, i.e. sentence–splitted, tokenized, morpho–syntactically tagged and lemmatized, consisting of 5,621 tokens and 214 sentences (hereafter referred to as *en_gold_EULaw*).

The source and target data are annotated according to the PTB[8] morpho–syntactic[9] and dependency tagsets.

3.2.1. Source vs Target Domain Corpora Annotation Criteria

The legal text contains pecularities regarding surface characteristics as well as dependency annotations that are hardly if at all present in the newspaper source domain data.

With regard to sentence splitting the same criteria were used as for Italian: in order to preserve the original structure of the law text, punctuation marks such as semicolon and colon that are followed by a carriage return are treated as sentence boundary markers. If no carriage return was present in the original text, the sentence was kept as is, thus resulting in some relatively long sentences. An example thereof is given in Figure 1. It will also serve as example to discuss some of the adopted annotation criteria. For instance, subsequent subordinate clauses without the main clause that are not present in the in–domain data, i.e. the subordinates introduced by *whereas* in our example. In this case, we chose to annotate the first instance of *whereas* as the *ROOT* node of the sentence and the second one as verbal modifier of the head of the preceding clause.

As will be shown further in Section 4., sentence length deviates considerably between the source and target domains. Another surface property of the target domain text that is different from the source domain is that the legal text contains a large amount of enumerations (lists, either hyphenated or enumberated with characters, numbers or roman numerals). In fact, one third (72 of 214) sentences in *en_gold_EULaw* are list items. Only very few of them (less

[3] A description of the part-of-speech (coarse– and fine–grained) tagsets and of the morpho–syntactic features can be found at http://poesix1.ilc.cnr.it/ISST-TANL-MStagset-web.pdf and at http://poesix1.ilc.cnr.it/ISST-TANL-MS_FEATStagset-web.pdf respectively.

[4] A description of the dependency tagset can be found at http://poesix1.ilc.cnr.it/ISST-TANL-DEPtagset-web.pdf

[5] http://medialab.di.unipi.it/wiki/SemaWiki

[6] http://www.cis.upenn.edu/∼treebank/

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42

[8] The head and dependency relation fields were converted using the algorithms described in (Johansson and Nugues, 2007).

[9] The fine grained part–of–speech are the gold standard part of speech tags from the WSJ, details of which can be found, http://bulba.sdsu.edu/jeanette/thesis/PennTags.html or http://www.cis.upenn.edu/∼treebank/

(13) Whereas Member States should be able to require that a prior consultation be undertaken by the party that intends to bring an action for an injunction, in order to give the defendant an opportunity to bring the contested infringement to an end ; whereas Member States should be able to require that this prior consultation take place jointly with an independent public body designated by those Member States ;

Figure 1: Example sentence from *en_gold_EULaw*.

than half a percent of the sentences in the PTB) are present in the source domain training data. We tried to treat the enumeration part consistently: if followed or surrounded by hyphens or parenthesis (like *(13)* in Figure 1), the list item marker was considered the head of the punctuation marks and attached to the head of the following sentence or phrase (as VMOD or NMOD). Moreover, since the PTB part–of–speech tags contains a respective "list item marker" tag (LS), the POS tags were tagged as such, accordingly. If the list item ended with a semicolon followed by a single conjunction (e.g. *; or*), it was attached as DEP (unclassified relation) to the head of he preceding clause. Further pecularities of the target domain (like the depth of embedded complement chains) are discussed in more detail in the following section.

3.3. Linguistic Preprocessing of Datasets

Both English and Italian datasets used for development and final testing have been morpho–syntactically tagged and lemmatized by a customized version of the pos–tagger described in Dell'Orletta (2009).

The manually annotated test sets were initially parsed by the DeSR parser (Attardi and Dell'Orletta, 2009), a state–of–the–art linear–time Shift–Reduce dependency parser, and were then manually revised by expert annotators, also on the basis of the extended annotation criteria reported in Sections 3.1.1. and 3.2.1. for Italian and English respectively.

4. Source vs Target Domain Data: Linguistic Features

In order to get evidence of the differences among the source and target domain data, the Italian and English distributed gold datasets have been monitored with respect to a number of different linguistic parameters. This allowed us to empirically *i)* define what we mean by *domain* and *ii)* to explain the drop of accuracy of general parsers on domain–specific texts and thus to motivate the need for developing domain adaptation strategies for reliably parsing of legal texts. As demonstrated by the results of the linguistic monitoring reported in the following sections, the two different legal language sub–varieties as well as the chemical and newswire texts each represent different classes of texts, henceforth generically referred to as *domains*, each characterized by specific linguistic features. The typology of features selected to reconstruct the linguistic profile characterizing each class of texts is organised into four main categories: raw text features, lexical features, morpho-syntactic and syntactic features. In what follows, we report and discuss the monitoring results obtained with respect to these different textual classes or domains.

Raw Text Features

The source domain and legal datasets for both Italian and English differ significantly in many aspects starting from the **average sentence length**, calculated as the average number of words per sentence[10] (see Figure 2). As Figure 2(a) shows, *it_NatRegLaw* contains the longest sentences with respect to all the other datasets. Interestingly, the sentence lengths of *it_gold_EULaw* and *en_gold_EULaw* sets are very close, i.e. 33.38 and 33.86 word–tokens respectively. This is mainly due to the fact that the two sets contain aligned sentences as well as to the nature of European legal texts, i.e. their being translations of an original unique text. It is also worth noting that the length of the sentences contained in *english_pchemtb* is closer to *english_ptb_train* and *english_ptb_test* than to *en_gold_EULaw*. This supports the hypothesis that chemical texts represent a different domain with respect to English European legislative texts.

Since, as claimed in the literature on measures of syntactic complexity (see below), a longer sentence is grammatically more complex than a shorter one, it can be argued that sentence length affects parsing accuracy. This is typically the case when such a feature is associated with long dependency links, as demonstrated by McDonald and Nivre (2007).

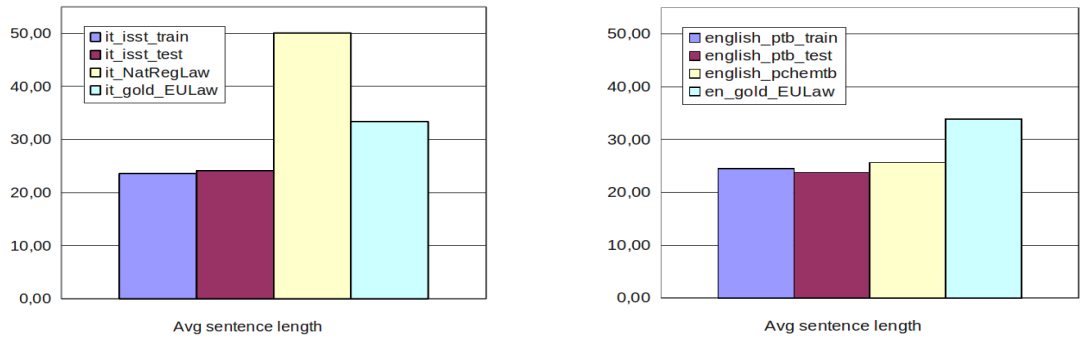
Lexical and Morpho–syntactic Features

Figure 3 reports the lexical overlap of the different corpora, calculated as the percentage of lexical items of *it_isst_train* and *english_ptb_train* also contained in the target domain test sets. First of all, it is worth noting that as far as *english_pchemtb* is concerned the percentage of newswire lexicon (0.60%) is lower than in *en_gold_EULaw* (0.86%). This allows highlighting a peculiarity of legal domain texts which contain a higher percentage of newswire lexicon than other domains. This finding is in line with what observed by Lease and Charniak (2005), who report the unknown word rate (expressed in terms of tokens) for various technical domains (e.g. biomedical abstracts, abstracts in the field of aereodynamics, etc.) which has been computed with respect to sections 2–21 of the WSJ.

As it can be seen in Figures 3(a) and 3(b), the lexicon specific to the legal domain is not extremely different from the one of the newswire domain. Interestingly, this holds true both for the Italian and English legal language used in texts enacted by the European Commission. This suggests that the main differences between newswire and legal texts are mostly concerned with the underlying syntactic structure. Nevertheless, a difference between the two considered Italian legal language sub–varieties exists: the percentage of newswire lexicon contained in *it_NatRegLaw* (0.81%) is lower than the one observed in *it_gold_EULaw* (0.88%).

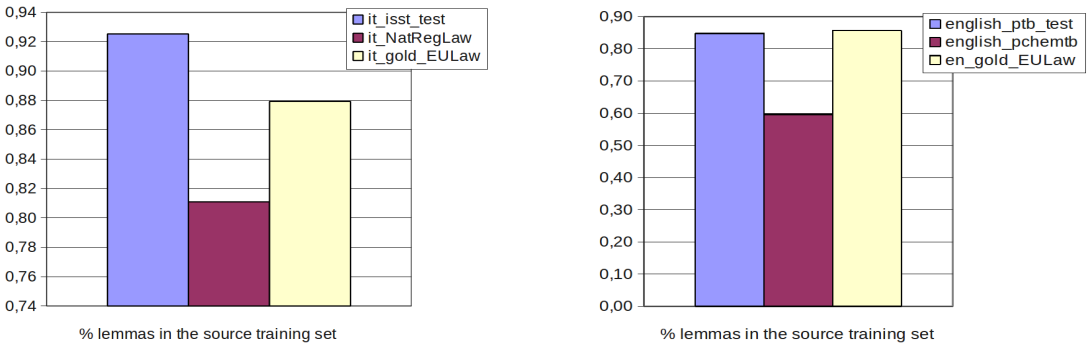
A last remark is in order here for what concerns the percentage of lexical items that the *it_isst_test* and *english_ptb_test* share with the corresponding training sets: the lexicon of the Italian test set turned out to be much more similar

[10]Note that sentence shorter than 5 word–tokens are excluded from the computation.



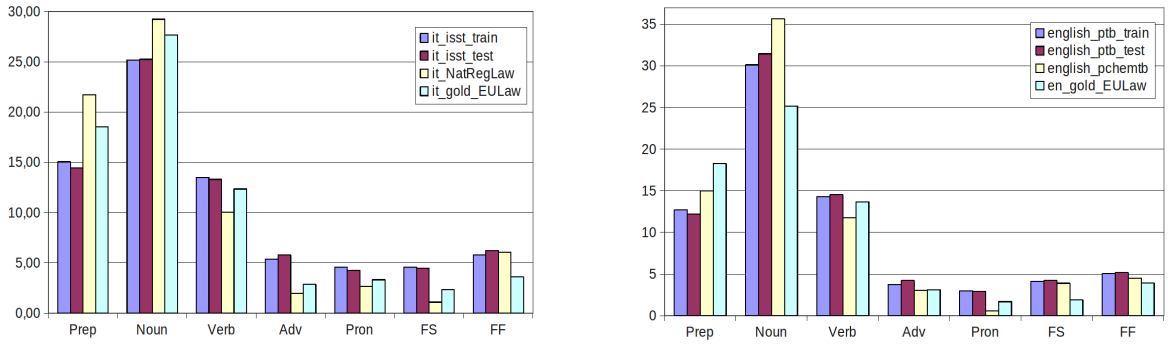
(a) Italian gold data (b) English gold data

Figure 2: Average sentence length in the Italian and English gold datasets.



(a) Italian gold data (b) English gold data

Figure 3: % of training set lemmas contained in the Italian and English gold datasets.



(a) Italian gold data (b) English gold data

Figure 4: Distribution of some of the main parts–of–speech in the Italian and English gold datasets.

(0.93%) to *it_isst_train* than the lexicon of *english_ptb_test* (0.85%) with respect to *english_ptb_train*. This follows from the strategy adopted for selecting the sentences contained in the test set: the sentences of *it_isst_test* have been randomly selected from the whole ISST–TANL corpus, while those in *english_ptb_test* have been taken from a section of the Penn Treebank different from the one included in *english_ptb_train*.

Let us focus now on the morpho–syntactic level. Figure 4 reports that different varieties of the legal language represented by *it_NatRegLaw*, *it_gold_EULaw* for Italian and *en_gold_EULaw* for English show a similar distribution of parts–of–speech: namely, they all have a higher percentage of prepositions (*Prep*) with respect to the ISST–TANL and PTB datasets, and a lower percentage of **verbs** (*Verb*), **adverbs** (*Adv*), **pronouns** (*Pron*), **punctuation marks**, i.e. full stops (*FS*) and commas (*FF*). These observed distributions can be taken as some of the main peculiar features of both Italian and English legal texts.

While the different distribution of punctuation marks can support the hypothesis of a sentence structure specific to legal texts, the high occurrence of prepositions can be strongly connected with their presence within long sequences of complements (see below for more details).

Surprisingly enough, the percentage distribution of **nouns** (*Noun*) is quite different across languages, i.e. in *it_gold_EULaw* and *en_gold_EULaw*. Similarly to *it_NatRegLaw*, the Italian European legal texts contain a higher percentage of nouns with respect to the ISST–TANL datasets. On the contrary, the occurrences of nouns in *en_gold_EULaw* are fewer than in the PTB data.

Syntactic Features

Major differences hold at the level of considered syntactic features, for which we observe a peculiar distribution which characterizes legal texts with respect to the source domain as well as to the other target domain datasets.

The first monitored syntactic feature is concerned with the **average depth of embedded complement 'chains'** governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers. Figures 5(a) and 5(b) show that both the Italian and English European legal texts are characterized by an average depth which is higher than the one observed in the ISST–TANL and PTB source domain datasets. This represents the syntactic counterpart of the peculiar distribution of prepositions observed in legal texts at the morpho–syntactic level (see above). Interestingly, the difference holding between the average depth of complement 'chains' occurring in *english_pchemtb* and the one observed with respect to the PTB dataset is less sharp than the difference between *en_gold_EULaw* and newswire PTB data. This demonstrates that the occurrence of deep embedded complement 'chains' appears to be a syntactic feature characterizing the legal domain with respect to newswire texts as well as to other domains. In Italian, this domain–specific feature appears to be more marked in the legal language sub–variety represented by *it_NatRegLaw*, which shows the deepest complement 'chains'.

A further distinguishing feature of legislative texts, still connected with the previous one, appears to be the different percentage distributions of embedded complement 'chains' by depth. As Figures 5(c) and 5(d) show, Italian and English legislative texts appear to have *i)* a lower occurrence of 'chains' including just one complement and *ii)* a higher percentage of deep complement 'chains' with respect to newswire data. Notably, *it_NatRegLaw* contains chains up to 9 embedded complements long.

It goes without saying that these two features can have a strong impact on the performances of parsers trained on the syntactic distributions of newswire texts.

The considered gold datasets have also been compared with respect to *i)* the **average length of dependency links**, measured in terms of the words occurring between the syntactic head and the dependent (with the exception of the punctuation marks), and *ii)* the **average depth of the whole parse tree**, calculated in terms of the longest path from the root of the dependency tree to some leaf. It has been chosen to monitor these two features since they both can be indicative of the structural complexity of a dependency structure. If on the parsing side McDonald and Nivre (2007) report that statistical parsers have a drop in accuracy when analyzing long distance dependencies, on the other hand Lin (1996) and Gibson (1998) claim that the syntactic complexity of sentences can be predicted with measures based on the length of dependency links, given the memory overhead imposed by very long distance dependencies. Parse tree depth is another feature reflecting sentence complexity as stated by, to mention only a few, Yngve (1960), Frazier (1985) and Gibson (1998).

As it can be seen in Figure 6, *i)* Italian and English legislative texts contain much longer (on average) dependency links than newswire texts and *ii)* the average height of *it_gold_EULaw* and *en_gold_EULaw* parse trees is higher than in the case of ISST–TANL and PTB. In addition, as it was previously pointed out, *it_NatRegLaw* texts appear to be syntactically more distant from newswire texts than European legal texts (see Figure 6(a)).

Finally, we compared source and target domain data with respect to the **arity of verbal predicates**, calculated as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers). A low arity value seems to be a distinctive feature of both Italian and English legal texts in comparison with newswire texts (see Figure 7). As Figure 7(a) shows, *it_NatRegLaw* contains verbal predicates characterized by the lowest arity. As suggested by Venturi (2011), this distinguishing feature of legal texts can be due to the frequent occurrence of verbal participial forms and of elliptical constructions.

5. Participation Results

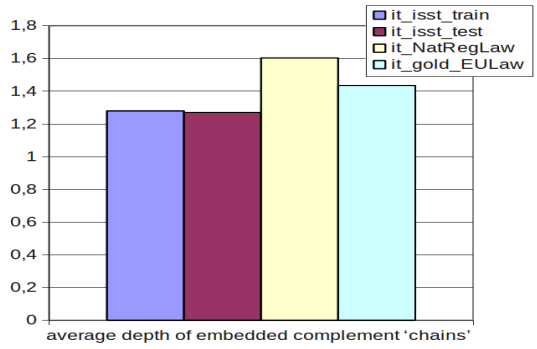
The participants to the shared task were three, namely **Attardi_et_al.** (University of Pisa, Italy), **Mazzei_Bosco** (University of Turin, Italy) and **Nisbeth_Søgaard** (University of Copenhagen, Denmark). Whereas the latter two teams participated only in the basic **Dependency Parsing** (DP) subtask for the Italian language, the first participant presented results for both languages and for both DP and **Domain Adaptation** (DA) subtasks.

5.1. Base Parsing Models

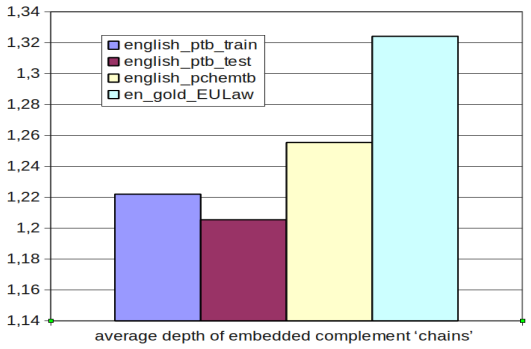
All participants adopted ensemble–based systems in which several base parsers produce dependency trees, which are then combined using different weighting functions (to weigth each dependency arc) and different combination algorithms.

Attardi_et_al. used a combination strategy exploiting the approximate linear time combination algorithm described by Attardi and Dell'Orletta (2009). The combined parsers are three different configurations of *DeSR* (Attardi, 2006), which is a Shift/Reduce deterministic transition–based parser that by using special rules is able to handle non–projective dependencies in linear time complexity. The configurations are: two versions differing with respect to the used learning algorithm (MultiLayer Perceptron (MLP) vs Support Vector Machine (SVM)) of the two stage Reverse Revision parser (i.e. a stacked righ-to-left parser that uses hints produced by a first pass left–to–right parser, Attardi and Dell'Orletta (2009)), and a right–to–left parser using an MLP classifier.

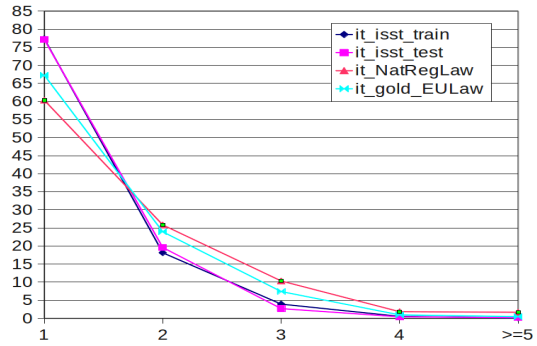
Mazzei_Bosco used a combination strategy based on a simple voting approach: for each word of the sentence the algorithm assigns the dependency head and dependency label more voted from the combined parsers and in the case



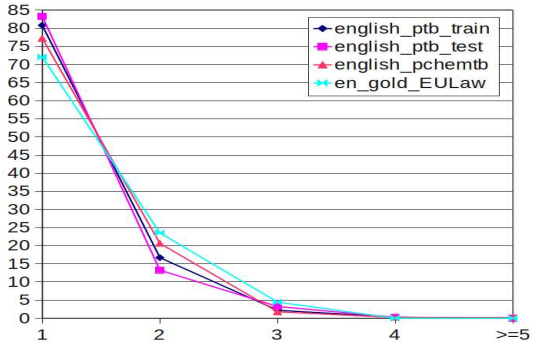
(a) Italian gold data



(b) English gold data

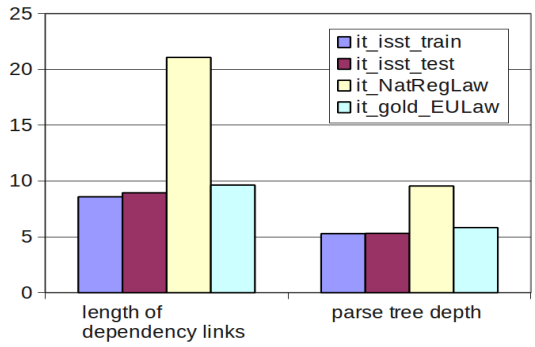


(c) Italian gold data

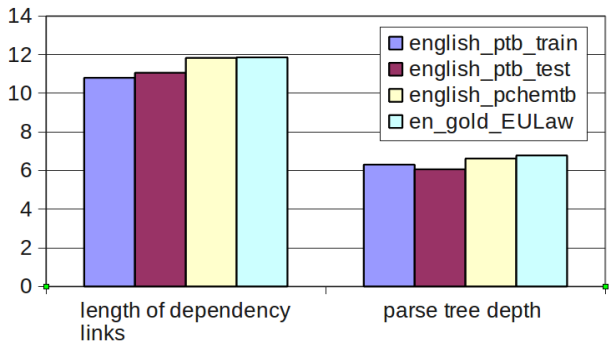


(d) English gold data

Figure 5: Average depth of embedded complement 'chains' (first row) and their distribution by depth (second row) in the Italian and English gold datasets.



(a) Italian gold data



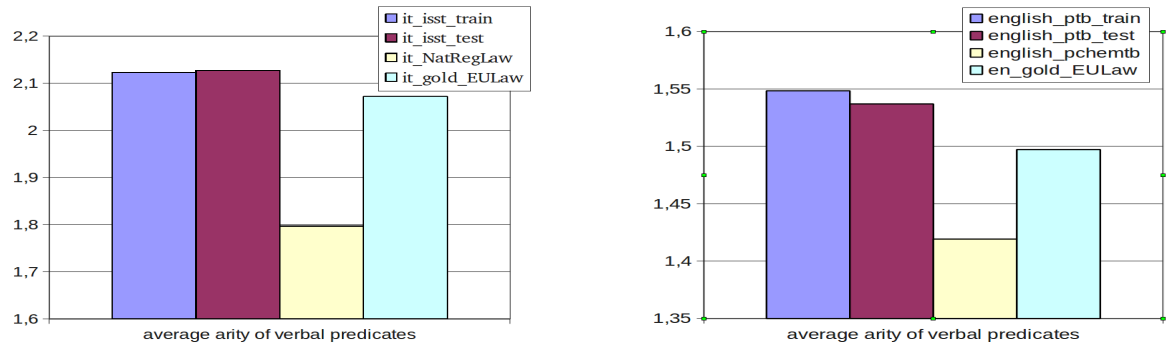
(b) English gold data

Figure 6: Length of dependency links and parse tree depth in the Italian and English gold datasets.

that each parser assigns a different dependency, the algorithm selects the dependency assigned by the best parser. Whenever in the resulted dependency structure there are cycles, the algorithm selects the tree produced by the best parser. Three different parsers are combined: *i)* left–to–right DeSR, using MLP as learning algorithm; *ii)* MaltParser (Nivre et al., 2006), a Shift/Reduce transition–based parser composed by a nondeterministic transition system for mapping sentences to dependency trees and a classifier that predicts the next transition for every possible system configuration (SVM was used as learning algorithm); *iii)* MateParser (Bohnet, 2010), an efficient implementation of the second order maximum spanning tree dependency parsing algorithm of Carreras (Carreras, 2007). The parser is trained using the margin infused relaxed algorithm (MIRA) (McDonald et al., 2005) and combined with a hash kernel (Shi et al., 2009).

Nisbeth_Søgaard adopted the combination strategy introduced by Sagae and Lavie (2009): using the analyses generated by the component parsers and a weighting function, a weighted directed graph is created where each word in the sentence is a node; finally, a maximum spanning tree algorithm is used to select the final analysis. To produce this combination they used MaltBlender software[11]. The ensemble system is based on several unoptimised parsers: *i)*

[11] w3.msi.vxu.se/users/jni/blend/



(a) Italian gold data

(b) English gold data

Figure 7: Average arity of verbal predicates in the Italian and English gold datasets.

ten instances of the MaltParser, one for each of the learning algorithms it provides; *ii)* one istance of the MateParser; *iii)* two istances (projective and non–projective) of MSTparser (McDonald et al., 2006), i.e. a graph–based parser which uses a maximum spanning tree algorithm for finding the highest scoring tree.

5.2. Results of the Dependency Parsing Subtask

Table 1 reports the results achieved by the participating systems on both the development in–domain test set (*it_isst_test*) and the out–domain test set (*it_gold_EULaw*) for the Dependency Parsing subtask for the Italian language. Unexpectedly, two out three participant parsing systems do not show a drop of accuracy when tested on the European legal texts. Interestingly, Mazzei_Bosco has an increment of 0.72 percentage points when their system was tested on the legal texts with respect to the newswire test.
This can be due to two main reasons. On the one hand, as already demonstrated by the results reported in (Sagae and Tsujii, 2007b), ensemble parsing systems are less affected by a drop of accuracy when tested on out–domain data in a domain adaptation scenario than single parsing systems, in particular when the types of parsing algorithms involved in the combination are different. On the other hand, as shown in Section 4., European legal texts are characterised by lexical, morpho–syntactic and syntactic features which make them not so distant from in–domain data.
On the contrary, the peculiar statistical distribution of monitored linguistic features in *national and regional Italian legal texts* (see Section 4.) can be seen as underlying the drop of accuracy of participant systems when tested on the out–domain development data provided (i.e. *it_NatRegLaw*), as reported in Table 2.

System	*it_isst_test*	*it_gold_EULaw*
Mazzei_Bosco	82.36	83.08
Attardi_et_al.	82.90	81.93
Nisbeth_Søgaard	81.43	81.58

Table 1: LAS for Dependency Parsing subtask for the Italian language.

Table 3 reports the Dependency Parsing results by

System	*it_NatRegLaw*
Mazzei_Bosco	75.88
Attardi_et_al.	74.03
Nisbeth_Søgaard	75.55

Table 2: LAS of participants on national and regional Italian legal texts.

Attardi_et_al. on both the development in–domain test set (*english_ptb_test*) and the out–domain test set (*en_gold_EULaw*) for the English language. Differently from Italian, for English we observe a noticeable drop of accuracy, of nearly 10 LAS percentage points. Different reasons can be seen as underlying this state of affairs. Among them, it is worth mentioning the occurence of syntactic structures specific to European legal texts and never occurring in the PTB test set for which new annotation criteria had to be defined (see Section 3.2.1.) and which can hardly be learned by a statistical parser trained on PTB. Moreover, the freer word order of Italian with respect to English can help explaining why statistical variations between the in– and out–domain texts might have a deeper impact on parser performances for English than for Italian: this is just an initial intuition which should be explored in more detail.

System	*english_ptb_test*	*en_gold_EULaw*
Attardi_et_al.	88.81	78.90

Table 3: LAS for Dependency Parsing subtask for the English language.

5.3. Results of the Domain Adaptation Subtask

For this subtask, Attardi_et_al. used a method based on active learning. They followed a two–step incremental process where each step generates a new training corpus including manually revised dependency–annotated sentences from the out–domain test unlabelled corpus. Each step can be summarised as follows: a) DeSR with MLP (Multi Layer Perceptron Algorithm) as learning algorithm is used to parse the unlabeled target corpus; b) perplexity mea-

sures based on the overall likelihood of the analysis of each sentence provided by DeSR are exploited to identify 100 sentences with the highest perplexity (*Lowest Likelihood*, LLK); and c) sentences selected during the previous step are manually revised and used to extend the training corpus in order to build a new parser model.
The new parser model was used to parse the target domain test set. For the last run they used the parser system described in section 5.1..

System	*it_isst_test*	*it_gold_EULaw*
Attardi_et_al.–run1		82.78
Attardi_et_al.–run2	82.05	83.52

Table 4: LAS for Domain Adaptation subtask for the Italian language.

System	*english_ptb_test*	*en_gold_EULaw*
Attardi_et_al.–run1	87.17	78.38

Table 5: LAS for Domain Adaptation subtask for the English language.

Tables 4 and 5 report the results achieved within the Domain Adaptation subtask for Italian and English respectively. *Attardi_et_al.–run1* and *Attardi_et_al.–run2* refer to the first and second step of the active learning process. For Italian, we can observe that the adopted domain adaptation strategy shows a significant parsing improvement: the parser shows a LAS improvement of 0.85 percentage points after the first added 100 sentences, and of 1.59 points after the second step. For English, the same DA strategy does not produce the same effect. After the active learning process, the parser has a drop of accuracy of 0.52 LAS percentage points. Among the reasons behind this drop there may be disalignments between gold annotations based on the new annotation criteria defined for dealing with legal texts (as discussed in Section 5.2.) and annotations performed by the annotators involved in the active learning process.
Tables 4 and 5 also report the results obtained for the in–domain development sets after the domain adaptation process: a small drop of accuracy can be observed. This is in line with what observed by McClosky et al. (2010) and Plank and van Noord (2011) who proved that parsers trained on the union of more than one different gold corpora taken from different domains achieve lower accuracy with respect to the same parsers trained on data belonging to a single target domain.

6. Conclusion

The SPLet 2012 shared task was the first competition on dependency parsing of legal texts. In this context, different parsing systems – all based on ensemble methods – have been tested against Italian and English legal data sets.
Different results have been achieved for the two languages. A significant drop in accuracy has been observed with respect to the English test set. Differently, for Italian two out of three participant systems showed no drop in accuracy against the final test set represented by European legal texts; however, the performance of all participant systems appear to significantly decrease when tested against texts belonging to the language sub–variety represented by national and regional legislative texts. This asymmetric behaviour of parsers can be explained by comparing the statistical distribution of linguistic features within in–domain training corpora and out–domain test sets. All participants used statistical parsers based on machine learning algorithms: this fact can help explaining why their performance decreases when parsing sentences characterized by features hardly or never occurring in the training set.
This prompts the need for domain adaptation strategies. In this shared task, only one system participated in the Domain Adaptation subtask by exploiting an active learning method which achieved good results for the European Italian legal texts. On the contrary, no improvement has been obtained for what concerns European English legal texts: this is very likely due to both language–specific peculiarities and annotation choices adopted to handle domain–specific syntactic constructions.
The SPLET 2012 Shared Task was successful in defining and analysing the stat–of–the–art performance of dependency parsing in the legal domain. The evaluation results of the final submissions for both subtasks from the participants are both promising and encouraging for the future of legal Information Extraction applications. Developed domain–specific annotated corpora together with descriptions of participant systems represent rich resources for finding directions for improvements. Last but not least, the experience of the shared task provides valuable input for facing further challenges specific to the domain.

7. Acknowledgments

The Shared Task Organizers would like to thank to the CoNLL 2007 Shared Task Organizers and the Linguistic Data Consortium for preparing and providing the in–domain and development English datasets. We also wish to acknowledge Tommaso Petrolito and Francesco Asaro for their contribution in annotating the Italian target domain data.

8. References

G. Attardi and F. Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL-HLT*.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Shared task CoNLL-X*, pages 166–170, New York City.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China.

C. Bosco and A. Mazzei. 2012. The evalita 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*, Rome, Italy.

C. Braun. 2003. Parsing german text for syntactico–semantic structures. In *Prospects and Advances in the Syntax/Semantics Interface, Proceedings of the Lorraine–Saarland Workshop*, Nancy, France.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, page 957961.

DANLP. 2010. *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing (2010)*. available at http://aclweb.org/anthology/W/W10/W10-2600.pdf.

F. Dell'Orletta, S. Marchi, S. Montemagni, G. Venturi, T. Agnoloni, and E. Francesconi. 2012. Domain adaptation for dependency parsing at evalita 2011. In *Working Notes of EVALITA 2011*, Rome, Italy.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of Evalita'09*, Reggio Emilia.

L. Frazier, 1985. *Syntactic complexity*. Cambridge University Press, Cambridge, UK.

B. Mortara Garavelli. 2001. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino, Einaudi.

E. Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

D. Gildea. 2001. Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 167–202, Pittsburgh, PA.

Richard Johansson and Pierre Nugues. 2007. *Extended Constituent-to-Dependency Conversion for English*. available at http://dspace.utlib.ee/dspace/bitstream/handle/10062/2560/reg-Johansson-10.pdf.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 58–69, Jeju Island, Korea.

D. Lin. 1996. On the structural complexity of natural language sentences. In *Proceedings of COLING 1996*, pages 729–733.

L.T. McCarty. 2007. Deep semantic interpretations of legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL2007)*, pages 217–224, Stanford, California.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California.

R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the EMNLP–CoNLL*, pages 122–131.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Association for Computational Linguistics*.

R. McDonald, K. Lerman, and F. Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL*, New York City.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: a data-driven parser-generator for dependency parsing. In *Proceedings of LREC-200*, pages 2216–2219.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel S., and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP–CoNLL*, pages 915–932.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1566–1576, Portland, Oregon.

Kenji Sagae and Alon Lavie. 2009. Parser combination by reparsing. In *Proceedings of HLT-NAACL*.

K. Sagae and J. Tsujii. 2007a. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1044–1050, Prague.

Kenji Sagae and Junichi Tsujii. 2007b. Dependency parsing and domain adaptation with lr models and parser ensemble. In *Proceedings of the EMNLP–CoNLL 2007*, pages 1044–1050.

Qinfeng Shi, JamesPetterson Gideon Dror, John Langford, Alex Smola, and S.V.N. Vishwanathan. 2009. Hash kernels for structured data. *Journal of Machine Learning*, 15(1):143–172.

G. Venturi. 2011. *Lingua e diritto: una prospettiva linguistico–computazionale*. Ph.D. Thesis, Università di Torino.

G. Venturi. 2012. Design and development of temis: a syntactically and semantically annotated corpus of italian legislative texts. In *Proceedings of the 4th Workshop Semantic Processing of Legal Texts (SPLeT 2012)*, Istanbul, Turkey.

S. Walter. 2009. Definition extraction from court decisions using computational linguistic technology. In G. Grewendorf and M. Rathert, editors, *Formal Linguistics and Law*, pages 183–224. Mouton de Gruyter.

V. H.A. Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society*, pages 444–466.

Active Learning for Domain Adaptation of Dependency Parsing on Legal Texts

Giuseppe Attardi, Daniele Sartiano and Maria Simi
Dipartimento di Informatica, Università di Pisa, Italy
e-mail: attardi@di.unipi.it, sartiano@di.unipi.it, simi@di.unipi.it

Abstract

Several techniques have been explored in the literature to achieve domain adaptation in parsing. In principle fully unsupervised methods would be preferable, but the evidence so far is that none of them is effective, except for one special case of self-training used within one step of a reranking constituency parser. For the task of domain adaptation of dependency parsing to legal text, we hence chose to use a semi-supervised technique (i.e. active learning) which has consistently proved effective in other types of domain adaptation. We report on how we used active learning, i.e. selection criteria, parameters used, to perform domain adaptation in two languages: Italian and English. The results are quite positive on Italian and less on English. We discuss possible explanations for this discrepancy.

Keywords: Dependency Parsing, Domain Adaptation, Active Learning.

1 Introduction

Statistical parsers have progressed significantly and currently they can reach high levels of accuracy when applied to sentences of the same genre as the training corpus on which they were trained. However accuracy may decrease significantly when they are applied to sentences from different domains. McClosky et al. (2010) report for instance a drop from an F1 score of 89% to 74% when applying a parser trained on the English Penn Treebank to sentences from the Genia corpus; Attardi et al. (2007) report a drop of Labeled Accuracy Score from 85.85% to 79.98 % when testing a dependency parser on a chemical domain.

It can be expected that a parser trained on a specific corpus be biased towards that corpus: even though certain aspects of language are general, at least terminology will differ across genres. Research on domain adaptations has been trying to close this gap, aiming at developing techniques that would allow a parser to learn features of a new domain, possibly using unsupervised methods since annotating corpora on each domain can be quite labour intensive.

2 Related Work

One way to reduce the effort of human annotation for each new domain is to use an *active learning* process.

Active learning is a semi-supervised machine learning technique in which the learner is allowed to choose the data from which it learns. An active learner generates *queries* for an *oracle* (e.g. a human annotator) to obtain labels for data instances selected from a larger set of unlabeled data. Active learning has been successfully applied in many modern machine learning problems where unlabeled data are abundant and easily obtained, but labelling is difficult, time-consuming, or expensive (Settles, 2010). In particular there is a growing interest in applying this technique to nearly all language technology tasks, as reported in a literature survey by Olsson (Olsson, 2009) and testified by the NAACL HLT 2009 Workshop on Active Learning for NLP (Ringger, Hertel & Tomanek, 2009).

Active learning has been applied successfully to parser domain adaptation on a question corpus by Atserias et al. (2010) and on legal texts in Evalita 2011 by Attardi et al. (2012).

Self-training is an unsupervised variant of active learning, where the learner itself performs the annotations for use in the next step of active learning on the basis of what it has learned so far. Self-training has been applied with mixed success to constituent parsing (Charniak 1977, Steedman et al., 2003) while slightly better results were obtained in combination with reranking (McClosky et al., 2006). These approaches only rely on information already available to the parser, for instance the POS of words surrounding the edges of each constituent.

Daumé III (2007) proposes a technique for learning the distinction between general and specific domain aspects. Pairing each feature with a "general" pseudo-domain (for capturing domain independent features) allows a classifier to directly model which features are domain-specific.

McClosky et al. (2010) address the issue of domain detection, in order to determine which one, among several parsers trained on different corpora, is the most appropriate to use.

3 Active Learning

For the task of domain adaptation to Legal Texts, we chose to use *active learning*, since it is the approach that gave most consistent results in our previous experiences.

The active learning process aims at reducing the human annotation effort, only asking for advice when the utility of the query is high. The primary question is therefore *query* formulation: how to choose which example (or examples) to try next.

There are many heuristics for choosing the examples: choosing examples where we don´t have data (Whitehead, 1991), where we perform poorly (Linden & Weber, 1993), where we have low confidence (Thrun & Möller, 1992; Donmez & Carbonell, 2008), where we expect it to change our model (Cohn et al., 1990), and where we previously found data that resulted in learning

(Schmidhuber & Storck, 1993).

Multi-classifier approaches use measures of disagreement among a committee of classifiers, obtained in different ways, as a measure of uncertainty (Freund et al., 1997).

A separate issue, which influences the speed and performance of the active learning process, is whether the learner should process a single instance or a batch of instances at each iteration. Adding one instance at a time slows the overall learning process down. If, on the other hand, a batch of instances is added, the learning progresses faster, but it becomes more difficult to find strategies for selecting a good batch. Metrics combining in various ways *informativeness* (inversely related to uncertainty), *representativeness* (related to density, computed with clustering techniques) and *diversity* (reducing repetitions) have been proposed to address this issue (Olsson, 2009). For example, in the context of statistical parsing, Tang et al. propose to cluster parsed sentences, represented as a series of parsing events, according to a similarity measure based on the Hamming distance[1]. A *representativeness* measure, based on the density of clusters, is then combined with a measure of uncertainty to form a selection criterion for sampling (Tang, Ruo & Roukos, 2002).

The optimal size of the batch is also a critical parameter, which needs to be tuned on the basis of the specific application.

Most of the empirical results in the published literature suggest that active learning works in practice, and *selective sampling* methods outperform *random sampling* (as typical in *passive* learning) in most learning task. This is often true even for simple query strategies, such as uncertainty sampling.

4 Training Collections

	Sentences	Avg. sent. Length	Tokens
English PTB	18,577	24.03	446,573
Italian ISST	3,275	21.85	71,568

Table 1. SPleT Corpora statistics.

Table 1 reports statistics of the corpora provided for the SPLeT Shared task at LREC 2012.

5 Approach

We addressed the issues stated in the introduction by means of *active learning*. Active learning is an iterative process where a learner is trained using an initial training set and then, the most representative examples, according to a suitable selection criterion, are selected from a non-annotated collection; these examples are manually annotated and added to the training corpus for the next iteration. After labeling every pattern re-compute interestedness of unlabeled points, choose the one with highest, label it, re-train, etc.

If the selection criterion is effective, a much smaller number of examples is sufficient to achieve the same level of accuracy than using normal supervised learning.

In classic AL the optimal size of data to add at each step is a single pattern. Adding more than one pattern at a time incurs in some loss of information, and as we add more and more in a batch we loose more and more information. In the extreme, if all the data were added at once, this would not be active learning.

For practical reasons we may want to add more than one pattern at a time, when, for example, re-training takes a long time and we do not want human annotators to wait. In this case, there is a trade-off between how long it takes to re-train and re-compute interestedness, how much can the annotators wait, and how much AL power we are willing to "loose". In practice, labelling several examples at a time in small batches is a good practice.

In our case we decided that a batch of 100 questions at a time is a quite conservative addition to a comparatively much larger training corpus.

6 Testing selection criteria

The first series of experiments aimed at observing the effect of different selection criteria, compared to random sampling.

6.1 Likelihood Estimates

We tested more sophisticated criteria to drive active learning based on likelihood estimates of a sentence parse. DesR is a transition-based parser (Attardi 2006), which uses a classifier to decide which action to perform to carry out parsing. The classifier computes a probability distribution for the possible actions to perform at each step. Given a parsed sentence, the probability of each parsing step is therefore available to compute different metrics by which to estimate the confidence of the parser in its own output. For example:

a. *Likelihood of a parse tree*, computed as the product of the probabilities of all the steps used in building the tree;
b. *Average probability* of the parsing steps in building the tree;

In our experiments we selected sentences according to three different ordering criteria:

1. *Lowest likelihood* of sentence parse tree (LLK): aims at preferring sentences that were judged more difficult, by considering the likelihood of the parse tree;
2. *Highest likelihood* of sentence parse tree (HLK): prefers sentences that were judged easier by the parser, by considering the likelihood of the parse tree;
3. *Lowest average probability* (LAP): selects sentences that were judged more difficult by computing the average probability of each parsing step;
4. *Lowest normalized likelihood* (LNL): takes into account the length of the parsed sentences by introducing a normalization factor (*likelihood*/*log*(*n*), where *n* is the number of tokens in the sentence).

7 Parsing Experiments

For parsing we used the latest version of DeSR (Attardi 2006), an efficient transition based dependency parser. The latest version, available on SourceForge (DeSR), provides a choice of classifiers, including MultiLayer Perceptron and SVM) as well as rich feature selection,

[1] The Hamming distance measures the number of substitutions required to turn a sequence into another.

including multiple feature combination.

For English we used a configuration similar to the one that gave the best results on the Penn English Treebank distributed for the CoNLL 2008 Shared Task. That version provided also annotations for lemmas and was bigger (39,279 sentences and 958,167 tokens) than the corpus provided for the current SPLeT task. DeSR achieves a Labeled Accuracy Score of 89.29 % on that collection.

The adapted feature configuration for English is reported below:

From single words
S_{0f} I_{0f} I_{1f} $prev(I_0)_f$ $lChild(I_0)_f$ $rChild(I_0)_f$ S_{2p} S_{1p} I_{0p} I_{1p} I_{2p} I_{3p} $prev(S_0)_p$ $next(I_1)_p$ $lChild(S_1)_p$ $lChild(I_0)_p$ S_{0c} I_{0c} I_{1c} $rChild(I_0)_c$ $rChild(rChild(I_0))_c$ $lChild(S_1)_d$ $lChild(I_0)_d$ $rChild(S_1)_d$
From word pairs
$S_{1c}I_{0c}$ $S_{0c}S_{1c}$ $S_{1c}I_{1c}$

Table 2. English feature model.

Features are extracted by the parser at each step from the current parser state, which consists of a triple $\langle S, I, A\rangle$, where S is the stack of previously analyzed tokens, I is the queue of (remaining) input tokens and A is the set of dependency arcs built so far.

We denote the tokens on the stack with S_0, S_1, etc., the front items from the input queue as I_0, I_1, etc., the head of x (if any) with $h(x)$, the leftmost and rightmost modifiers of x (if any) with $lChild(x)$ and $rChild(x)$, respectively, the token preceding x (if any) with $prev(x)$ and the one following x as $next(x)$. Attributes of a token are denoted by subscripts: f is the form, l is the lemma, p is the POS tag, c is the coarse POS tag, d is the dependency tag, m is the morphology.

The feature model for Italian is adapted from the one we used in the Domain Adaptation task at Evalita 2011. The parser trained for Evalita 2011 achieved a LAS of 82.09 %.

From single words
S_{1l} S_{0l} I_{0l} I_{1l} I_{2l} I_{3l} $prev(I_0)_l$ $lChild(S_0)_l$ $lChild(I_0)_l$ $rChild(S_0)_l$ $rChild(I_0)_l$ S_{2p} S_{1p} I_{0p} I_{1p} I_{2p} I_{3p} $next(S_0)_p$ $lChild(S_1)_p$ $lChild(S_0)_p$ $rChild(S_0)_p$ $rChild(I_0)_p$ S_{0c} I_{0c} I_{1c} $rChild(I_0)_c$ $rChild(rChild(I_0))_c$ S_{0m} I_{0m} S_{1m} $lChild(S_0)_d$ $lChild(I0)_d$ $rChild(S_1)_d$
From word pairs
$S_{1c}I_{0c}$ $S_{0c}S_{1c}$ $I_{0c}I_{1c}$ $S_{1c}I_{1c}$ $S_{1c}I_{2c}$ $S_{1c}I_{3c}$ $rChild(S_0)_cI_{0c}$

Table 3. Italian feature model.

8 Active Learning Experiments

As criterion for selecting sentences for Active Learning we used the *lowest likelihood* (LLK), as well as filtering sentences by length, criteria that turned out as the most effective in our previous experiences at domain adaptation for questions (LREC 2010) and in domain adaptation for Italian (Evalita 2011).

According to the Active Learning paradigm, at each step a new parser is trained on the corpus produced in the previous iteration. So after each step, the adaptation corpus, less those sentences added to the training corpus, is parsed again with the new model and all sentences ranked, according to *lowest likelihood*, for selecting a new batch of sentences to annotate and add to the training corpus for the next active learning step.

We submitted a run to the Basic Task and two runs to the Domain adaptation for Italian, corresponding to two steps of active learning.

For the Basic Task we used a parser combination, exploiting the technique by Attardi and Dell'Orletta (2009), consisting of a parser using an MLP classifier in a stacked reverse revision combination, a second MLP parser running in reverse and a third parser using an SVM classifier in stacked reverse revision combination.

The first step of Active Learning, used 100 sentences of length greater than 8 and less than 20, selected as the first 100 having a LogLikelihood score below -14. The latter value was chosen empirically to discard very poor sentences, for examples those including TAB or '|' characters or similar, apparently coming from tables.

The second step of Active Learning, added 100 more sentences of length between 8 and 30 with LogLikelihood below -23.

The parser used in the Domain Adaptation task was also a combination of the same type as in the Basic Task.

The Labeled Accuracy Score (LAS), achieved by these runs are reported in the Table below. For comparison, we report also the LAS on the ISST Test set of both the parser trained on the ISST training set and the one extended with Active Learnng.

Train Set	ISST Test Set	it_EULaw
ISST	82.90 %	81.93 %
+ AL1		82.78 %
+ AL1 + AL2	82.05 %	83.52 %

Table 4: Italian Results

The results show that the drop in LAS on the it_EULaw test set is minor and just the addition of 100 sentences in the first step of Active Learning is sufficient to close the gap.

The second step of Active Learning further improves the score and does not significantly worsen the accuracy on the base test set.

Train Set	PTB Test Set	en_EULaw
PTB CoNLL 2007	88.81%	78.90 %
+ AL1	87.17%	78.38 %
+ AL1 fixed	87.29%	80.83 %

Table 5: English Results

The results for English are less encouraging: the drop of accuracy from the original domain to the legal domain is significant, almost 10 percent points. Besides, active learning does not improve at all. This is partly due to problems in the annotation of the legal test set that we report below and that led us to repeat the run, which led to the improvement shown in the last line of the Table.

9 Annotation Issues

While for Italian the need for adaptation to the new

domain was minimal (our best parser reported only a drop of less than 1% LAS going from the ISST source test set to the it.EULaw test set), for English the drop in performance was more significant (nearly 10 percent points of LAS) suggesting that a more radical adaptation to the new domain was needed.

Unfortunately the active learning strategy did not work out as expected, resulting in a small additional drop in performance (-0.52%) after the addition of 100 manually revised sentences from the target domain.

These results were disappointing and did not match our expectations, arising from previous experience with active learning. Therefore we tried to understand the reasons of this behavior and came out with the following observations.

We noticed some disagreements between the conventions used by the annotators, namely our annotator supporting the active learning strategy on one hand and the annotators of the gold standard on the other hand; this mismatch was amplified by the fact that many sentences from the target domain were involved, resulting in a big impact on accuracy.

We report two major cases of disagreement, both very typical of the legal domain, as attested by their frequency in the test set:

1. The way the leading numbers (or letters) introducing items in numbered lists were annotated;
2. Constructions involving "*whereas*".

An example of the first case is the sentence:

(c) the meal plan:

The gold annotator always marked the opening and closing parenthesis as dependent of the enclosed numbering, consistently with how most balanced punctuations are dealt in the PTB corpus; the AL annotator instead chose as reference standard the annotation style used in PTB for sentences like:

2) Encourage long-term occupancy by forgiving ...

where the closed parenthesis is marked as dependent of the root of the sentence (*occupancy*). As a consequence the AL annotator consistently annotated every case of surrounding parenthesis as dependent of the head of the sentence.

In order to estimate the impact of this choice on the final parsing outcome, we rerun the active learning step after fixing the annotation of numbered sentences according to the gold convention. The accuracy increased to 80.83 % as reported in Table 5 in the row labelled + AL1 fixed.

The second case of annotation disagreement was of a different nature and is more related to an incoherent annotation in the gold test set. In 19 cases out of 40 occurrences of "*whereas*" in the gold test set, "*whereas*" appears as root of the sentence, while in the remaining 21 "*whereas*" is annotated as a VMOD of the verb, as the AL annotator did.

In the following phrase, both type of annotations occur within the same sentence:

Whereas products age in the course of time, higher safety standards are developed and the state of science and technology progresses; whereas, therefore, it would not be reasonable to make the producer liable for an unlimited period for the defectiveness of his product; whereas, therefore, liability should expire after a reasonable length of time, without prejudice to claims pending at law;

The first "*Whereas*" is annotated as ROOT, while the second as VMOD dependent of "*would*" and the third of "*should*".

The parser annotates all these cases consistently, and in our view, correctly, as dependents of the corresponding verb. Each of these annotations causes a penalty of at least 3 undue errors to the parser ("*whereas*", verb, and final punctuation going to a different root).

While some inter-annotator disagreement is to be expected, and certainly there are more of them, these two cases alone occur in a large percentage of the sentences (94 out of 227), and justify the relatively small drop in performance of the first step of adaptation. The parser had a chance to learn how to adjust to the new domain, but was misled by "wrong" indications.

In order to partially assess the accuracy drop due to these cases, we measured the accuracy of the parser output on the test set, dropping the sentences starting with "*Whereas*" or with numbering. The LAS increased from 78.40% to 82.62%. This however reduced the size of the test set from 5621 tokens to 2531.

10 Conclusions

We have been exploring the issue of domain adaptation for dependency parsing for several years. At CoNLL 2007 (Attardi et al., 2007) we used effectively the approach of Tree Revision (Attardi and Ciaramita, 2007). We tried this approach on the present task, but we did not achieve as good results as with active learning, possibly because we are using now a parser with greater accuracy and the technique is less capable of detecting recurring errors.

We have later attempted the approach of *Self Training*, which is similar to Active Learning, except that it selects sentences to add to the training set chosen among those that the parser itself produces, from unannotated data in the target domain. The results have been quite disappointing, like those of other studies that failed to show a benefit to parsing from self-training (Charniak, 1997; Steedman et al., 2003). An obvious reason might be that the sentences which hare added, are either indeed well parsed sentences, and therefore the parser has nothing to learn from them, or are incorrectly parsed ones, and hence they will confuse the training algorithm. Only McClosky et al. (2006) were able to achieve some improvement by self-training, but just in the special case of the first stage of a constituent reranking parser. Error analysis showed that improvements were correlated with the length of the sentence and the number of conjunctions but not with the number of unknown words in the sentence.

In a further attempt we tried to provide the parser with suggestions about which words might be related and hence be possible candidates for being related in a dependency. This information could be obtained from large sets of unannotated documents using measures like Pointwise Mutual Information. Unfortunately not even this approach turned out successful.

Finally we attempted using transductive SVM (Miceli Barone & Attardi, 2012) in order to exploit large amounts

of unannotated data in training. This approach is however computationally very expensive and produces only minor improvements.

In order to learn, one must supply new knowledge to the learning algorithm: active learning works because of the extra knowledge that is provided through newly human annotated data.

We applied active learning successfully to parsing domain adaptation tasks on adapting to parse questions (Atserias et al., 2010) and to legal texts both in Evalita 2011 and at the LREC 2012 Workshop on Semantic Processing of Legal Texts.

Our conclusion is that currently no other approach is more effective in dependency parsing domain adaptation than active learning. The approach is semi-automated, since it still requires the intervention of a human annotator. However, since we are offering an automated way to select the sentences to annotate and only a small number of these is sufficient to achieve adequate accuracy improvements, the technique remains practically viable.

11 Acknowledgments

This work has been supported with insignificant funding from project PRIN by the Italian ministry MIUR.

12 References

Atserias, J., Attardi, G., Simi, M., Zaragoza, H. (2010). Active Learning for Building a Corpus of Questions for Parsing. *Proc. of LREC 2010*, Malta.

Attardi, G. (2006). Experiments with a Multilanguage non-projective dependency parser. In *Proc. of the Tenth CoNLL*.

Attardi, G., Ciaramita, M.. (2007). Tree revision learning for dependency parsing, *Proceedings of HLT-NAACL 2007*, Rochester.

Attardi, G., Chanev, A., Ciaramita, M., Dell'Orletta, F. and Simi, M. (2007). Multilingual Dependency Parsing and Domain Adaptation using DeSR. *Proceedings the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

Attardi G., Dell'Orletta F. Reverse Revision and Linear Tree Combination for Dependency Parsing. *Proc. of NAACL HLT*, 2009.

Atardi, G., Simi, M., Zanelli, A. (2012). Parser Domain Adaptation by Active Learning. *Proc of Evalita 2011*. *LNCS*, Spinger-Verlag (forthcoming).

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proc. AAAI*, 598–603.

Cohn, D., Atlas, L. & Ladner, R. (1990). Training Connectionist Networks with Queries and Selective Sampling. In D. Touretzky (Ed.) *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann.

Daumé III, Hal. (2007). Frustratingly easy domain adaptation. In *Proceedings of ACL 2007*, Prague, Czech Republic.

DeSR. Dependency Shift Reduce Parser, http://desr.sourceforge.net/.

Donmez, P. & Carbonell, J. G. (2008). Optimizing estimated loss reduction for active sampling in rank learning. *International Conference on Machine Learning*, ICML.

Freund, Y., Seung, S., Shamir, E. & Tishby, N. (1997). Selective sampling using the query-by-committee algorithm, *Machine Learning*, 28, pp. 133–168.

Linden, A. & Weber, F. (1993). Implementing inner drive by competence reflection. In H. Roitblat et al. (Eds.), *Proc. 2nd Int. Conf. on Simulation of Adaptive Behavior*, MIT Press, Cambridge.

McClosky, D., Charniak, E. & Johnson, M. (2006). Reranking and self-training for parser adaptation. *Proc. of the 44th ACL*, 337–344.

McClosky, D., Charniak, E. & Johnson, M. (2010). Automatic Domain Adapatation for Parsing. *Proc. of NA ACL – HLT 2010 Conference*, Los Angeles, CA.

Miceli Barone, A.V. & Attardi, G. (2012). Dependency Parsing domain adaptation using transductive SVM. *Proc. of Workshop Robus-Unsup*.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing, Swedish Institute of Computer Science Technical Report, April 17.

Ringger, E., Hertel, R. & Tomanek, K. (Eds.) (2009). *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*.

Settles, B. (2010). Active Learning Literature Survey, Computer Science Technical Report 1648, University of Wisconsin-Madison, Jan. 26.

Schmidhuber, J. & Storck, J. (1993). Reinforcement driven information acquisition in nondeterministic environments. Tech. Report, Fakultat Informatik, Technische Universitat München.

Steedman, M., et al. (2003). CLSP WS-02 Final Report: Semi-Supervised Training for Statistical Parsing. Technical report, Johns Hopkins University.

Surdeanu, M., Ciaramita, M. & Zaragoza, H. (2008). Learning to Rank Answers on Large Online QA Collections. *Proc. of the 46th Annual Meeting of the ACL*.

Tang, M., Luo, X. & Roukos, S. (2002). Active Learning for Natural Language Parsing, *Proc. of the 40th Annual Meeting of the ACL*, 120–127.

Thrun, S. & Möller, K. (1992). Active exploration in dynamic environments. In J. Moody et al. (Eds.) *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann.

Whitehead, S. (1991). A study of cooperative mechanisms for faster reinforcement learning. TR-365, Dept. of Computer Science, Rochester Univ., Rochester, NY.

Zhang, Y. & Nivre, J. (2011). Transition-based dependency parsing with rich non-local features. *Proc. of the 49th Annual Meeting of the ACL*.

Simple Parser Combination

Alessandro Mazzei, Cristina Bosco

Dipartimento di Informatica, Università degli studi di Torino
Corso Svizzera 185, 10149, Torino, Italy
mazzei@di.unito.it, bosco@di.unito.it

Abstract

This paper presents an ensemble system for dependency parsing: three parsers are separately trained and combined by means of a majority vote. The three parsers are (1) the MATE parser [http://code.google.com/p/mate-tools/], (2) the DeSR parser [http://sites.google.com/site/desrparser/], and (3) the MALT parser [http://maltparser.org/]. The MATE, that was never used before on Italian language, drastically outperforms the other parsers in the SPLeT shared task. Nonetheless, a simple voting combination further improves its performances.

Keywords: ensemble parsing, MATE parser, DeSR parser, MALT parser

1. Introduction

In last few years parsing community devoted great attention to dependency formalisms, and today dependency parsing can be seen as the first step in many applicative NLP systems (Kübler et al., 2009). Larger dependency treebanks and more sophisticated parsing algorithms allowed improved performances of dependency parsers for many languages (Nivre et al., 2007; Hajič et al., 2009).

Indeed, dependency parsing performances constantly increased for Italian. As reported in the Evalita evaluation campaigns specific for NLP systems for Italian (EVALITA 2011 Organization Comitee, 2012), the best scores for Italian dependency parsing (expressed in Labelled Attachment Score, LAS) was 86.94% in 2007, 88.73% in 2009, and 91.23% in 2011 (Bosco and Mazzei, 2012). These results were obtained by using the Turin University Treebank, a dependency treebank for Italian (Bosco and Lombardo, 2004) (see the Section 4.). However, statistical dependency parsing seems to be still improved. On the one hand, new promising specific algorithms for learning and classification are emerging; on the other hand researchers are applying universal machine learning techniques to this specific task. Some are trying to use lager sets of syntactic features (e.g. (McDonald and Pereira, 2006; Carreras, 2007)), while others are trying to apply general techniques *to combine* together the results of various parsers (Zeman and Žabokrtskỳ, 2005; Sagae and Lavie, 2006; Hall et al., 2007; Attardi and dell'Orletta, 2009; Surdeanu and Manning, 2010; Lavelli, 2012).

Our system in the SPLeT competition follows both these mentioned directions. We employ three state of art statistical parsers, which use sophisticated parsing algorithms and advanced feature sets. The three parsers are (1) the MATE parser (Bohnet, 2010), (2) the DeSR parser (Attardi, 2006), (3) the MALT parser (Nivre et al., 2006). Moreover, in our system we combine these three parsers by using two very simple voting algorithms (Breiman, 1996; Zeman and Žabokrtskỳ, 2005). We decided to apply an "out of box" approach, i.e. we apply each parser with its standard configurations for learning and classification.

In the next Sections we first give a short description of the three parsers (Section 2.), then we describe our approach for ensemble parsing (Section 3.) and we report the results of our experiments (Section 4.), before to conclude the paper (Section 5.).

2. The three parsers

In this Section we give a brief description of the three parsers applied in our experiments, i.e. MATE, DeSR and MALT parser.

The MATE parser (Bohnet, 2009; Bohnet, 2010) is a development of the algorithms described in (Carreras, 2007; Johansson and Nugues, 2008). It basically adopts the second order maximum spanning tree dependency parsing algorithm. In particular, Bohnet exploits *hash kernel*, a new parallel parsing and feature extraction algorithm that improves the accuracy as well as the parsing speed (Bohnet, 2010). The MATE performances on English and German, which are 90.14% and 87.64% respectively (LAS), posed this parser at the state of art for these languages (Hajič et al., 2009; Bohnet, 2010; Anders et al., 2010).

The DeSR parser (Attardi, 2006) is a transition (shift-reduce) dependency parser similar to (Yamada and Matsumoto, 2003). It builds dependency structures by scanning input sentences in left-to-right and/or right-to-left direction. For each step, the parser learns from the annotated dependencies if to perform a shift or to create a dependency between two adjacent tokens. DeSR can use different set of rules and includes additional rules to handle non-projective dependencies. The parser can choose among several learning algorithms (e.g Multi Layer Perceptron, Simple Vector Machine), providing user-defined feature models. In our experiments we adopted for DeSR the Multi Layer Perceptron algorithm, which is the same configuration that the parser exploited when it won the Evalita 2009 competition.

The MALT parser (Nivre et al., 2006) implements the transition-based approach to dependency parsing too. In particular MALT has two components: (1) a (non-deterministic) transition system that maps sentences to dependency trees; (2) a classifier, that predicts the next transition for every possible system configuration. MALT performs a greedy deterministic search into the transition system guided by the classifier. In this way, it is possible to perform parsing in linear time for projective dependency

trees and quadratic time for arbitrary (non-projective) trees (Nivre, 2008). MALT has several built-in transition systems, but in our experiments we adopted just the standard "Nivre arc-eager" system, that builds structure incrementally from left to right. Moreover, we use the standard classifier provided by MALT, i.e. the SVM (Simple Vector Machine) basic classifier on the standard "NivreEager" feature model.
In our knowledge this is the fist work that experimented the MATE parser on Italian, while DeSR and MALT parsers have been used in many occasions on Italian (e.g. (Lavelli, 2012; Attardi et al., 2012)), reaching the best results in several contests.

3. The combination algorithms

In order to combine the three parsers we used two very simple algorithms, COM1 and COM2, both implemented in PERL programming language. These algorithms have been previously experimented in (Zeman and Žabokrtskỳ, 2005) and in (Surdeanu and Manning, 2010).
The main idea of the COM1 algorithm is to do a democratic voting among the parsers. For each word[1] of the sentence, the dependency (parent and edge label) assigned to the word by each parser is compared: if at least two parsers assign the same dependency, the COM1 algorithm selects that dependency. In the case that each parser assigns a different dependency to the word, the algorithm selects the dependency assigned by the "best parser", that in our experiments on development set was the MATE parser (see below). As noted by (Zeman and Žabokrtskỳ, 2005), that uses the name *voting* for COM1, this is the most logical decision if it is possible to identify a priori the "best parser", in contrast with the more democratic random choice.
The COM2 algorithm is a variation of the COM1. COM1 is a single word combination algorithm that does not consider the whole dependency structure. This means that incorrect dependency trees can be produced by the COM1 algorithm: cycles and several roots can *corrupts* the "treeness" of the structure. The solution that we adopt in the COM2 algorithm is very simple: if the tree produced by the COM1 algorithm for a sentence is corrupted, then it is selected as dependency structure for that sentence the tree produced by the "best parser" . Again, in accord (Zeman and Žabokrtskỳ, 2005), that uses the name *switching* for COM2, this is the most logical decision since MATE is without doubts the best parser on development score.

4. Experimental Results

We used two machines for experiments. A powerful Linux workstation, equipped with 16 cores, processors 2GHz, and 128 GB ram has been used for the MATE parser, so that the average time for learning is 8 hours. Another Linux workstation equipped with a a single processor 1GHz, and 2 GB ram has been used for learning of the DeSR and MALT parsers, that usually required a couple of hours, and for testing that required several minutes for MATE parser and few minutes for MALT and DeSR parsers. MALT and

[1] In this paper we use the term *word* in a general sense, as synonym of *token*.

```
FOREACH sentence
  FOREACH word IN sentence
    IF [L-DeSR(word)==L-MALT(word)]
      L-COM1(word):=L-DeSR(word)
    ELSE
      L-COM1(word):=L-MATE(word)
```

Table 1: The combination algorithm COM1, that correspond to the *voting* algorithm reported in (Zeman and Žabokrtskỳ, 2005)

```
FOREACH sentence
  FOREACH word IN sentence
    IF [L-DeSR(word)==L-MALT(word)]
      L-COM2(word):=L-DeSR(word)
    ELSE
      L-COM2(WORD):=L-MATE(WORD)
  IF [!CORRECT(TREE-COM2)]
    T-COM2(sentence):=T-MATE(sentence)
```

Table 2: The combination algorithm COM2, that correspond to the *switching* algorithm reported in (Zeman and Žabokrtskỳ, 2005)

DeSR parsers accept as input the CONLL-07 format, that is the format provided by the SPLeT organizers. In contrast MATE accept the CONLL-09 format: simple conversions scripts have been implemented to manage this difference.
In the first experiment, in order to evaluate the "best parser" in the COM1 and COM2 algorithms, we used the ISST training (file: $it\_isst\_train.splet$, $71,568$ words, $3,275$ sentences) as learning set and the ISST development (file: $it\_isst\_test.splet$, $5,165$ words, 231 sentences) as development set. The second row in Table 3 shows the results of the three parsers in this first experiment. MATE parser outperforms the DeSR and MALT parsers: in particular, MATE does $\sim 3\%$ better than DeSR and $\sim 5\%$ better than MALT. On the basis of this result, we decided to use MATE as our "best parser" in the combination algorithms (cf. Section 3.). COM1 and COM2 reach the score of 82.54% and 82.36% respectively, and so both combination algorithms improve the performances of the MATE parser close to the 0.5%.
In the second experiment, we use the whole ISST as learning set (files: $it\_isst\_train.splet$ and $it\_isst\_test.splet$, total $76,733$ words, $3,506$ sentences) and we use the blind file provided by the organizers as test set (file: $it\_EULaw\_test\_blind.splet$, $5,662$ words, 240 sentences, European Directives Laws). The first row in Table 3 shows the results of the three parsers in this second experiment: the value 83.08%, produced by the COM2 algorithm, is the final result of our participation to the SPLeT shared task.[2] Note that there is a $\sim 0.1\%$ difference between the COM1 and COM2 results: similar to (Zeman and Žabokrtskỳ,

[2] A previous value of 84.95% was computed on the basis of two misunderstandings: (1) the NatReg set was added to the learning set and (2) the COM1 algorithm was used (instead of the COM2) since it was not assumed the tree-structure constraint.

	MATE	DeSR	MALT	COM1	COM2	Blended$_{W_2}$	Blended$_{W_3}$	Blended$_{W_4}$
TestSet	82.57	78.68	77.98	83.20	**83.08**	82.23	83.15	83.24
DevSet	81.92	78.99	77.04	82.54	82.36	81.45	82.54	82.63
NatReg	75.76	70.66	70.33	76.28	75.88	74.78	76.07	75.97
Evalita11	89.07	86.26	80.76	89.19	89.16	88.03	89.19	89.19

Table 3: The performances (LAS score) of the three parsers, their simple combination (COM1 and COM2), their blended combination (Blended$_{W_2}$, Blended$_{W_3}$, Blended$_{W_4}$) on the SPLeT test set, development set, Regional laws set and on the Evalita test.

2005; Surdeanu and Manning, 2010) we have 10 corrupted trees in the test set, i.e. $\sim 4\%$ of the total (240 sentences). In Table 4 we detailed the results of the three parsers in the second experiment on the basis of their agreement. When the three parsers agree on the same dependency (Table 4, first row), this happens on $\sim 72\%$ of the words, they have a very high LAS score, i.e. 95.6%. Moreover, DeSR and MALT parsers do better of the MATE parser only when they agree on the same dependency (Table 4, second row). The inspection of the other rows in Table 4 shows that COM1 algorithms has the best possible performance w.r.t. the voting strategy. In other words, COM1 selects all the parser combinations that correspond to higher value of LAS score (cf. the discussion on *minority dependencies* in (Surdeanu and Manning, 2010)).

In the third experiment, we again use the whole ISST as learning set (files: $it\_isst\_train.splet$ and $it\_isst\_test.splet$, total $76,733$ words, $3,506$ sentences), but we use the NatReg file provided by the organizers as test set (file: $it\_NatRegLaw\_test\_blind.splet$, $5,194$ words, 119 sentences, Regional Laws of Piedmont Region). The third row in Table 3 shows the results of the three parsers in this third experiment: in this case we have 75.88% for COM2 algorithm. This lower result can be advocated to the different nature of the domain. It is interesting to note that in this experiment MALT and DeSR parsers give similar results ($\sim 70\%$), while the MATE parser still outperforms them by $\sim 5\%$.

Finally, we performed a fourth experiment on totally different learning and test sets, by using a different Italian Treebank with a different PoS tag set and a different dependency format. We used the Evalita 2011 Development Set as learning set (file: $evalita2011\_train.conll$, $93,987$ words, $3,452$ sentences; balanced corpus of newspapers, laws, wikipedia) and we use the Evalita 2011 test as test set (file: $evalita2011\_test.conll$, $7,836$ words, 300 sentences; balanced corpus), that are produced by using the Turin University Treebank (Bosco and Mazzei, 2012). The fourth row in Table 3 shows the results of the three parsers in this third experiment: in this case we have 89.16% for COM2 algorithm[3]. It is interesting to note that the improvement of the COM2 algorithm w.r.t. with respect to the MATE parser is only $\sim 0.1\%$. In Table 5 we detailed the results of the three parsers in this fourth experiment on

[3]This score is the third w.r.t. to Evalita 2011 dependency parsing shared task, where the Parsit Parser achieved the best score (91.23%) the DeSR parser achieved the second best score (89.88%).

Scores					Frequency
MATE 95.6	==	DeSR	==	MALT	**71.99**
MATE 30.7	!=	**DeSR 45.8**	==	**MALT**	**4.20**
MATE	==	**DeSR 67.2**	!=	MALT 14.4	**7.70**
MATE	==	**MALT 59.1**	!=	DeSR 20.0	**8.21**
MATE 31.1	!=	DeSR 14.5	!=	MALT 16.3	**7.89**

Table 4: The detailed performances (LAS score) of the three parsers and their simple combination on the SPLeT blind set, i.e. corresponding to the first row of the Table 3.

the basis of their agreement. Again, when the three parsers agree on the same dependency (Table 5, first row), this happens on $\sim 78\%$ of the words, they have a very high LAS score, i.e. 96.6%. In contrast with the second experiment, here we have a not relevant improvement when DeSR and MALT parser do better of the MATE parser, i.e. only when they agree on the same dependency (Table 5, second row). In other words, on the SPLeT test set the COM1 (and COM2[4]) algorithm do much better than MATE since DeSR and MALT parsers have a good performance (45.8% vs. 30.7%) when they do not agree with the MATE parser: this is not true for the Evalita11 experiment, where DeSR and MALT have 38.8% while the MATE has 35.2%.

Combining versus Re-parsing

Since COM1 can produce corrupted dependency trees, as in (Zeman and Žabokrtskỳ, 2005) we used the COM2 algorithm, that checks the correctness of the tree and, in case of tree-corruption, returns the dependency structure produced by the "best parser" of the ensemble. We hypothesize that this strategy can produce good results in our system since one of the parser of the ensemble drastically outperforms the others. However, a general solution to the tree-corruption problem has been proposed: the re–parsing strategy (Sagae and Lavie, 2006; Hall et al., 2007; Attardi and dell'Orletta, 2009). In re–parsing, a new (not corrupted) dependency tree is produced by taking into account the tree produced by each parser of the ensemble: (Attardi and dell'Orletta, 2009) proposed a approximate top-down algorithm that starts by selecting the highest-scoring root

[4]In the fourth experiment there are 8 corrupted trees.

Scores					Frequency
MATE 96.6	==	DeSR	==	MALT	**78.39**
MATE 35.2	!=	**DeSR 38.8**	==	**MALT**	**3.38**
MATE	==	**DeSR 82.0**	!=	MALT 7.2	**9.17**
MATE	==	**MALT 63.3**	!=	DeSR 19.6	**4.27**
MATE 40.7	!=	DeSR 18.4	!=	MALT 7.9	**4.78**

Table 5: The detailed performances (LAS score) of the three parsers and their combination on the Evalita 2011 test set, i.e. corresponding to the fourth row of the Table 3.

node, then the highest-scoring children and so on; (Sagae and Lavie, 2006; Hall et al., 2007) apply a two-steps algorithm: (1) create a graph funding all the structures produced by the parser on the ensemble, and (2) extract the most probable dependency spanning tree from this graph.
(Surdeanu and Manning, 2010) provided experimental evidence that re–parsing algorithms are a better choice for practical ensemble parsing in out-domains: in order to test this hypothesis we performed a number of experiment by using the "MaltBlender" tool (Hall et al., 2007). In Table 3, the columns **Blended**$_{W_2}$, **Blended**$_{W_3}$, **Blended**$_{W_4}$ report the application of the algorithm described in (Hall et al., 2007). There are three weighting strategies: the results of the three parsers are equally weighted (W_2); the three parsers are weighted according to the total labeled accuracy on a held-out development set (W_3); the parsers are weighted according to labeled accuracy per coarse grained PoS tag on a held-out development set (W_4).
For the first, the second and the third experiments (Table 4, first second and third row), the held-out development set is the SPLeT development set; for the fourth experiment (Table 4, fourth row), the held-out development set is the Evalita 2011 test set. Three evidences seems to emerge from this last experiment: (1) the re–parsing strategies always performs slightly better than COM2 algorithms but not always better than COM1 algorithm; (2) there is no winning weighting strategy for re–parsing; (3) it does not seem that blending performs better out-domain than in-domain.

5. Conclusions

In this paper we described our parsing system for the participation to the SPLeT 2012 Shared Task, and two main issues arise by our contribution. The first issue is that the MATE parser has very good performance on Italian ISST treebank, both in domain and out domain, reaching very good scores; similar results have been obtained on the Turin University Treebank. The second issue is that very simple combination algorithms, as well as more complex blending algorithms, can furthermore improve performance also in situations where a parser outperforms the other ones.
In future research we plan to repeat our experiments on larger set of parsers. In particular, on the basis of the consideration that "diversity" is an important value in ensemble parsing, we want to experiment the possibility to combine together statistical parsers with rule based parsers, e.g. (Lesmo, 2012).

Acknowledgements

We want to thank Alessia Visconti and Francesca Cordero for their valuable (human and machine) time. Moreover we like to thank Felice Dell'Orletta for the suggestion to use MaltBlender in the analysis of the results.

6. References

Björkelund Anders, Bohnet Bernd, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

Giuseppe Attardi and Felice dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *HLT-NAACL*, pages 261–264.

Giuseppe Attardi, Maria Simi, and Andrea Zanelli. 2012. Tuning DeSR for the Evalita 2011 Dependency Parsing. In *Working Notes of EVALITA 2011*. CELCT a r.l. ISSN 2240-5186.

Giuseppe Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 166–170, New York City, June. Association for Computational Linguistics.

Bernd Bohnet. 2009. Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 67–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Cristina Bosco and Vincenzo Lombardo. 2004. Dependency and relational structure in treebank annotation. In *Proceedings of the COLING'04 workshop on Recent Advances in Dependency Grammar*, Geneve, Switzerland.

Cristina Bosco and Alessandro Mazzei. 2012. The evalita 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*. CELCT a r.l. ISSN 2240-5186.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 957–961.

EVALITA 2011 Organization Comitee. 2012. *Working Notes of EVALITA 2011*. CELCT a r.l.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared

task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johan Hall, Jens Nilsson, Joakim Nivre, Gülsen Eryigit, Beáta Megyesi, Mattias Nilsson, and Markus Saers. 2007. Single malt or blended? a study in multilingual parser optimization. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 933–939.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 183–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sandra Kübler, Ryan T. McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Alberto Lavelli. 2012. An Ensemble Model for the EVALITA 2011 Dependency Parsing Task. In *Working Notes of EVALITA 2011*. CELCT a r.l. ISSN 2240-5186.

Leonardo Lesmo. 2012. The Turin University Parser at Evalita 2011. In *Working Notes of EVALITA 2011*. CELCT a r.l. ISSN 2240-5186.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006))*, volume 6, pages 81–88.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: a data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*, volume 2216-2219.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computional Linguistics*, 34(4):513–553, December.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.

Mihai Surdeanu and D. Christopher Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *NAACL*. The Association for Computational Linguistics.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*, volume 3.

D. Zeman and Z. Žabokrtskỳ. 2005. Improving parsing accuracy by combining diverse dependency parsers. In *International Workshop on Parsing Technologies. Vancouver, Canada*, pages 171–178. Association for Computational Linguistics.

Parser combination under sample bias

Niklas Nisbeth, Anders Søgaard

University of Copenhagen
Njalsgade 140, DK-2300 Copenhagen
niklas@nisbeth.dk, soegaard@hum.ku.dk

Abstract

Combining several parsers through voting is known to improve parsing performance and robustness in supervised parsing. The intuition behind our shared task contribution to SPLeT 2012 is that voting is particularly useful when labeled data is biased, e.g. in domain adaptation.

Keywords: voting, sample bias, dependency parsing

1. Introduction

Voting is known to improve the performance and robustness of classifiers (Lin et al., 2003). If we have three base classifiers c_1, c_2, c_3, we simply predict the class of $\mathbf{x}$ is

$$\text{mode}\{c_1(\mathbf{x}), c_2(\mathbf{x}), c_3(\mathbf{x})\}$$

Rather than a simple plurality vote we can also predict a weighted vote

$$\begin{aligned}\arg\max_{y\in\mathcal{Y}} & \, w(c_1)||c_1(\mathbf{x}) = y|| \\ & + w(c_2)||c_2(\mathbf{x}) = y|| \\ & + w(c_3)||c_3(\mathbf{x}) = y||\end{aligned}$$

where w is some weighting function assigning a weight to each classifier, e.g. an estimate of its accuracy. The submitted results used a weighting function where the weights also depend on the data points:

$$\begin{aligned}\arg\max_{y\in\mathcal{Y}} & \, w(c_1, \mathbf{x})||c_1(\mathbf{x}) = y|| \\ & + w(c_2, \mathbf{x})||c_2(\mathbf{x}) = y|| \\ & + w(c_3, \mathbf{x})||c_3(\mathbf{x}) = y||\end{aligned}$$

In particular the weight function assigns estimated accuracies on subsets of the data.

Dependency parsing is a structured prediction task where the hidden variables are trees with complex internal structure. Two different parsers will rarely predict the same tree for a sentence of reasonable length, so base parsers cannot vote on entire trees. Sagae and Lavie (2006) propose to vote on dependencies in a round of greedy head selection. The votes are then used to build a weight matrix which is given as input to a minimum spanning tree algorithm singling out the dependency tree that maximizes the number of possibly weighted votes. We use the same reparsing scheme here.

While voting is known to correct the structural biases of parsers, the intuition behind our shared task contribution is that different structural biases also make parsers vulnerable to different kinds of bias in the labeled data. So voting should, if the necessary conditions for voting are satisfied, lead to larger error reductions when labeled data is biased. When using binned voting we bin weights on the head word's part of speech.

Parser	Algorithm	LAS
maltparser	2planar	78.28
	planar	78.13
	covnonproj	77.76
	nivreeager	77.69
	stacklazy	77.15
	stackproj	76.85
	covproj	76.67
	stackeager	76.37
	nivrestandard	75.56
mate-tools	nonproj	**80.98**
mstparser	nonproj	78.28
	proj	76.37

Table 1: Base parsers

2. Experiments

We use the data provided by the shared task organizers and use in-domain development data to estimate accuracies for weighted and binned voting. We experimented with 12 parsing algorithms from three publicly available parsers.[1][2][3] The performance of our base parsers is given in Table 1.

We experimented with the three voting schemes mentioned above, referred to below as plurality voting, weighted voting, and binned voting. We submitted a binned vote using all 12 dependency parsers as our ensemble. Results are presented in Table 2 with the submitted result in italics. We observe that using only the five best parsers on development data gives a slightly better result. Finally, we report the performance of the best ensemble that can be built from the 12 parsers.

The plot in Figure 1 shows how performance increases by ensemble size up to ensemble size 7. Dotted lines represent the macro-average performance of all possible ensembles of a given size. The line "maj-av" is the average performance of ensembles using plurality voting; and so on. The straight dotted lines are the baseline and our submitted results.

[1] http://maltparser.org
[2] http://code.google.com/p/mate-tools/
[3] http://sourceforge.net/projects/mstparser/

Ensemble	plurality	weighted	binned
all	81.58	81.76	*81.58*
5-devbest	-	-	81.76
best	82.32	82.32	**82.48**

Table 2: Voting

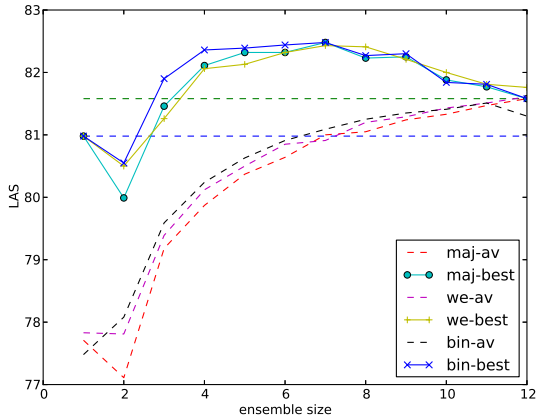


Figure 1: Performance by ensemble size

Somewhat surprisingly weighted voting does not outperform plurality voting, but binned voting is, as also observed in Sagae and Lavie (2006), superior to both plurality voting and weighted voting. Interestingly, however, the differences between the voting schemes level out as ensemble size increases.

The error reduction of 7.9% is a little less than expected,[4] but comparable to the result reported in Fishel and Nivre (2009) when using voting to parse the Brown corpus with parsers trained on PTB. Related work in parser combination includes Hall et al. (2007), Surdeanu and Manning (2010), Haffari et al. (2011).

3. References

Mark Fishel and Joakim Nivre. 2009. Voting and stacking in data-driven dependency parsing. In *NODALIDA*.

Gholamreza Haffari, Marzieh Razavi, and Anoop Sarkar. 2011. An ensemble model that combines syntactic and semantic clustering for discriminative dependency parsing. In *ACL*.

Johan Hall, Jens Nilsson, Joakim Nivre, Gülsen Eryigit, Beáta Megyesi, Mattias Nilsson, and Markus Saers. 2007. Single malt or blended? In *CoNLL*.

Xiaofan Lin, Sherif Yacoub, and Steven Simske. 2003. Performance analysis of pattern classifier combination by plurality voting. *Pattern Recognition Letters*, 24:1959–1969.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *HLT-NAACL*.

Mihai Surdeanu and Christopher Manning. 2010. Ensemble models for dependency parsing: cheap and good? In *NAACL*.

[4] Sagae and Lavie (2006) report an error reduction of 19%.