Workshop Programme

9:00-9:30

Welcome and introduction

9:30-10:00

Giulia Venturi. *Parsing Legal Texts. A Contrastive Study with a View to Knowledge Management Applications*

10:00-10:30

Karel Pala, Pavel Rychly and Pavel Smerk. *Automatic Identification of Legal Terms in Czech Law Texts*

10:30-11:00

Coffee break

11:00-11:30

Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Alessandro Mazzei, Daniele Radicioni and Piercarlo Rossi. *Multilevel Legal Ontologies*

11:30-12:00

Eneldo Loza Mencia and Johannes Fürnkranz. *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain*

12:00-12:30

Rafal Rzepka, Hideyuki Shibuki, Yasutomo Kimura, Keiichi Takamaru, Masafumi Matsuhara and Koji Murakami. *Judicial Precedents Processing Project for Supporting Japanese Lay Judge System*

12:30-13:00

Yasuhiro Ogawa, Kazuhiro Imai and Katsuhiko Toyama. *Evaluation Metrics for Consistent Translation of Japanese Legal Sentences*

13:00-13:30

Emmanuel Chieze, Atefeh Farzindar and Guy Lapalme. *Automatic Summarization and Information Extraction from Canadian Immigration Decisions*

13:30

Workshop end

Workshop Organisers

Simonetta Montemagni (Istituto di Linguistica Computazionale, CNR, Italy)
Daniela Tiscornia (Institute of Legal Information Theory and Techniques, CNR, Italy)
Enrico Francesconi (Institute of Legal Information Theory and Techniques, CNR, Italy)
Wim Peters (Natural Language Processing Research Group, University of Sheffield, UK)

Programme Committee

Danièle Bourcier (Humboldt Universität, Berlin, Germany)
Paul Bourgine (CREA, Ecole Polytechnique, Paris, France)
Joost Breuker (Leibniz Center for Law , University of Amsterdam, Netherlands)
Pompeu Casanovas (Institut de Dret i Tecnologia, UAB, Barcelona, Spain)
Alessandro Lenci (Dipartimento di Linguistica, Università di Pisa, Italy)
Leonardo Lesmo (Dipartimento di Informatica, Università di Torino, Torino, Italy)
Manfred Pinkal (Department of Computational Linguistics, Saarland University, Germany)
Vito Pirrelli (Istituto di Linguistica Computazionale of CNR, Pisa, Italy)
Paulo Quaresma (Universidade de Évora, Portugal)
Erich Schweighofer (Universität Wien, Rechtswissenschaftliche Fakultät, Wien, Austria)
Tom van Engers (Leibniz Center for Law, University of Amsterdam, Netherlands)
Maria A. Wimmer (Institute for Information Systems, Koblenz University, Germany)
Radboud Winkels (Leibniz Center for Law, University of Amsterdam, Netherlands)

Table of Contents

Preface

Giulia Venturi. *Parsing Legal Texts. A Contrastive Study with a View to Knowledge Management Applications* 1

Karel Pala, Pavel Rychly and Pavel Smerk. *Automatic Identification of Legal Terms in Czech Law Texts* 11

Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Alessandro Mazzei, Daniele Radicioni and Piercarlo Rossi. *Multilevel Legal Ontologies* 16

Eneldo Loza Mencia and Johannes Fürnkranz. *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain* 23

Rafal Rzepka, Hideyuki Shibuki, Yasutomo Kimura, Keiichi Takamaru, Masafumi Matsuhara and Koji Murakami. *Judicial Precedents Processing Project for Supporting Japanese Lay Judge System* 33

Yasuhiro Ogawa, Kazuhiro Imai and Katsuhiko Toyama. *Evaluation Metrics for Consistent Translation of Japanese Legal Sentences* 42

Emmanuel Chieze, Atefeh Farzindar and Guy Lapalme. *Automatic Summarization and Information Extraction from Canadian Immigration Decisions* 51

Author Index

Gianmaria Ajani, 16
Guido Boella, 16
Emmanuel Chieze, 51
Atefeh Farzindar, 51
Johannes Fürnkranz, 23
Kazuhiro Imai, 42
Yasutomo Kimura, 33
Guy Lapalme, 51
Leonardo Lesmo, 16
Eneldo Loza Mencia, 23
Masafumi Matsuhara, 33
Alessandro Mazzei, 16
Koji Murakami, 33
Yasuhiro Ogawa, 42
Karel Pala, 11
Daniele Radicioni, 16
Piercarlo Rossi, 16
Pavel Rychly, 11
Rafal Rzepka, 33
Hideyuki Shibuki, 33
Pavel Smerk, 11
Keiichi Takamaru, 33
Katsuhiko Toyama, 42
Giulia Venturi, 1

Preface

This volume contains the papers accepted for presentation at the LREC 2008 Workshop on “Semantic processing of legal texts”, held in Marrakech, Morocco on the 27th of May 2008.

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 addressing aspects such as automated legal reasoning and argumentation, semantic and cross-language legal information retrieval, document classification, legal drafting, legal knowledge discovery and extraction. Many efforts have also been devoted to the construction of legal ontologies and their application to the law domain.

A number of different Workshops and Conferences have been organised on these topics in the framework of the artificial intelligence and law community: among them, the ICAIL (International Conference on Artificial Intelligence and Law) and the Jurix (International Conference on Legal Knowledge and Information Systems) conferences, different workshops on Legal Ontologies (LEGONT) or on Legal Ontologies and Artificial Intelligence Techniques (LOAIT). The availability of lexical resources to enforce semantic interoperability among legal information is an emerging topic within the workshops held by the Legal XML Community (LegalXML workshops and Legal XML Summer school). In all these events, the topics of language resources and human language technologies are receiving increasing attention.

On the other hand, little attention has been paid to the legal domain within the computational linguistics community besides a few and isolated contributions and/or projects devoted to the processing of legal texts.

In this situation, we thought that time was ripe for offering a workshop on Language Resources (LRs) and Human Language Technologies (HLTs) in the legal domain, in which the two communities could 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. In the call for papers we solicited papers focussed on the topic of automatically extracting relevant information out of legal texts and of providing a structured organisation of extracted knowledge, and in particular on the crucial role played by language resources and human language technologies.

The response to the call for papers and the quality of the submitted papers mark this as a promising field which combines legal informatics and natural language processing in innovative and productive ways. If on the one hand this is a very encouraging fact, on the other hand we feel that much research and development remains to be carried out and that such an event will be beneficial to both communities, with the legal artificial intelligence community gaining insight on state-of-the-art linguistic technologies, tools and resources, and the computational linguists taking advantage of the large and often multilingual legal resources – corpora as well as lexicons and ontologies - for training and evaluation of current NLP technologies and tools.

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 2008 Organising Committee that made this workshop possible.

Workshop Chairs

Simonetta Montemagni
Daniela Tiscornia
Enrico Francesconi
Wim Peters

Parsing legal texts. A contrastive study with a view to Knowledge Management Applications

Giulia Venturi
Istituto di Linguistica Computazionale, CNR, Pisa
Via G. Moruzzi, 1 56124 Pisa, Italy
giulia.venturi@ilc.cnr.it

Abstract

Because of its inherent complexity, little attention has been paid so far to the linguistic analysis of law texts. Generally referred *structural obscurities of linguistic realisations specific of legal language* have been hampering efforts at using NLP techniques to law text semantic processing. In this work we investigate what these legal sublanguage peculiarities are; our analysis is mostly focused on those syntactic features which can make legal language different from ordinary language. For this purpose we used NLP techniques, exploiting a shallow parsing approach, i.e. chunking. We carried out a comparative study between a corpus of Italian law texts and an Italian general language corpus. Moreover, some of the legal language syntactic peculiarities detected were observed to be shared by a corpus of English law texts. Our main aim is to highlight the importance of investigating the linguistic structure underlying law texts in order to endow knowledge management applications (e.g. semantic annotation and event extraction) with "linguistic intelligence".

1. Introduction and motivation

The problem of automatically extracting semantic information out of the enormous and steadily growing amount of electronic text data is becoming more and more pressing. To overcome this problem, various technologies for information management systems have been explored within the Natural Language Processing (NLP) and AI communities. Two promising lines of research are represented by the investigation and development of information extraction technologies for a) ontology learning from document collections, and b) semantic mark-up of texts. In both cases, the text-to-knowledge process needs to rely on advanced NLP techniques.

The current state-of-the-art attests that many research activities related to ontology learning from text have been concerned with the (semi)automated ontology building using NLP techniques. According to Buitelaar et al. (2005), the ontology development task can be seen as a stratification process which provides a hierarchical cascade of knowledge layers. That is, the whole process can be organised in a "layer cake" of increasingly complex subtasks. Going from the extraction of terminological information (i.e. term extraction) to more abstract generalisations such as concept and relation extraction, many efforts have been frustrated by the difficulties inherent in natural language processing.

Similar observations are held in the case of semantic annotation which also needs to rely on advanced NLP techniques (see Reeve and Han, 2005 for a survey of the state-of-the-art). In particular, the annotation of inter-entity relational information (going from relations such as *place_of*, *author_of* to the specific events in which entities take part), which is becoming more and more a quite crucial task, requires the explicit representation of the linguistic micro-structure of a text.

The situation in the legal field is made even more pressing by the fact that laws are invariably conveyed through natural language. One of the main problems is caused by the nature of legal language and from its peculiarities as a technical language. This undermines the understandability of law texts, often producing difficulties in effective access to legislative resources by both legal experts and citizens.

In this paper, we will focus on the importance of endowing knowledge management systems with "linguistic intelligence" as well as of relying on advanced NLP techniques for semantic annotation purposes. We suggest that linguistic peculiarities of text need to be investigated and taken into account for any higher-level content analysis.

To our knowledge, little attention has been paid to the linguistic analysis of legal documents within the legal AI community. In particular, few attempts have been devoted to investigate legal language peculiarities which can be responsible for hampering law text semantic processing. Exceptions are the cases of Van Gog and Van Engers (2001), Lame (2005), Saias and Quaresma (2005), Walter and Pinkal (2006) and Moens et al. (2007). Indeed, this analysis is far from being straightforward because of the well-known complexities of legal language.

This study contributes to shed light on the main legal language features with a specific view to knowledge management applications. For this purpose, a battery of NLP tools was exploited and results obtained on corpora of Italian law texts were analysed. Legal language peculiarities were identified by comparing those results with the ones obtained on an Italian reference corpus. For this comparative analysis, the output of the same battery of NLP tools was inspected. Finally, a case study was carried out in order to investigate whether and to what extent legal language peculiarities are specific of the Italian case or if they can be found in other languages as well. Thus, a contrastive linguistic analysis was carried out comparing Italian and English legal corpora.

This paper is organised as follows: section 3 describes the NLP techniques exploited for the analysis of law texts. Section 4 shows the results of morpho-lexical and mainly syntactic analysis carried out on Italian legal corpora. Results of a comparative analysis with an Italian general language corpus are also reported. Then, section 5 is focused on the English law texts analysis. Section 6 outlines the main syntactic peculiarities shared by Italian and English law texts. Finally, in section 7 we suggest how a domain-specific linguistic analysis could be exploited for further semantic analyses as well as for future, more complex, knowledge management applications.

2. NLP analysis of legal texts

Our analysis of law texts focused on morphological and syntactic levels. For what concerns syntax, we focussed on *chunking*, the shallow syntactic parsing technique which segments sentences into an unstructured sequence of syntactically organised texts units called *chunks* (Abney, 1991). Abney (1991) in his study demonstrated how chunking has been proven to be highly versatile to produce reliable syntactic annotations of texts. The purpose of traditional full-parsing is to associate to each sentence a fully specified recursive structure, in order to identify the proper syntagmatic composition, as well as the relations of functional dependency among the identified constituents. On the contrary, chunking is a process of non-recursive segmentation of text. The resulting analysis is flat and unambiguous: only those relations which can be identified with certainty have been found out. Accordingly, some of the ambiguous grammatical dependencies (e.g. noun sequences, adjective conjunction, prepositional phrase attachments, etc.) are left underspecified and unresolved. This makes chunking highly suitable to syntactically annotate different types of texts, both written and spoken, and to analyse corrupted or fragmentary linguistic inputs. As long as "parse incompleteness" is reinterpreted as "parse underspecification", failures due to lexical gaps, particularly complex syntactic constructions, as well as ill-formed inputs, etc. are minimised. This allows chunking to be a starting point for parsing a language as complex as the legal one.

Within the nowadays Natural Language Processing community there is a spread consensus for exploiting this shallow parsing scheme as a reliable approach towards a robust syntactic analysis. Among others, Lenci et al. (2001) claim that «while full parsing is an extremely costly task for most existing systems since it needs huge amounts of linguistic knowledge to work properly, NLP systems can resort to a shallower level of syntactic description, which although underspecified, still provides enough syntactic information as the basis for higher-level processing tasks».

Moreover, although it might seem that full parsing should be preferred for an adequate processing of texts, the general tendency among the information extraction community is towards a shallow parsing approach to syntactic analysis. Recently, Bartolini et al. (2004b) have shown in their work the main advantages in taking chunked syntactic structure as the basis on which further stages of legal text processing operate. It has been reported there that «chunked representations can profitably be used as the starting point for partial functional analyses, aimed at reconstructing the range of dependency relations within the law paragraph text that are instrumental for the semantic annotation of text». The major potential for text chunking lies in the fact that «chunking does not "balk" at the domain-specific constructions that do not follow general grammar rules; rather it actually carries on parsing, while leaving behind an ill-formed chunk unspecified for its category». Walter and Pinkal (2006) also showed how making use of syntactic underspecification can help in dealing with the problem of the ambiguity of legal texts and, accordingly, how this can improve the quality of an information access and an ontology learning task within the legal domain.

We proceed with the idea that a shallow parsing approach can help to provide enough detailed linguistic information even for syntactically complex texts such as legal ones. Taking into account an unambiguous syntactic representation such the one provided by chunking, we would like to figure out the main syntactic peculiarities specific of legal language as well as the inherent complexities which make legal text processing a quite critical task.

3. The analysis of Italian legal texts

3.1. The NLP tools

AnIta (Bartolini et al., 2004) is the parsing system used for the analysis of Italian law texts. It is constituted by a pipeline of NLP tools also including a chunking module, CHUG-IT (Federici et al, 1996). In CHUG-IT chunking is carried out through a finite state automaton which takes as input a morpho-syntactically tagged text. Under Federici et al.'s interpretation, «a *chunk* is a textual unit of adjacent word tokens: accordingly, discontinuous chunks are not allowed. Word tokens internal to a chunk share the property of being mutually linked through those dependency chains which can be identified unambiguously with no recourse to lexical information other than part of speech and lemma». To be more concrete, a sentence such as *Le stesse disposizioni si applicano ad un prodotto importato* "The same provisions are applied to an imported product" will be chunked as follows:

A. [Le stesse disposizioni] *The same provisions*
B. [si applicano] *are applied*
C. [ad un prodotto] *to a product*
D. [importato] *imported*

"Le stesse disposizioni si applicano ad un prodotto importato" ("The same provisions are applied to an imported product")

```
[[CC:N_C][DET:LO#RD][PREMODIF:STESSO#A][POTGOV:DISPOSIZ
IONE#S]]
[[CC:FV_C][CLIT:SI#PQ][POTGOV:APPLICARE#V]]
[[CC:P_C][PREP:AD#E][DET:UN#RI][POTGOV:PRODOTTO#S]]
[[CC:ADJPART_C][POTGOV:IMPORTARE#V@IMPORTATO#A]]
```

Figure 1: CHUG-IT output

A sample output of CHUG-IT is given in Figure 1, where it can be noted that each chunk contains information about its type (e.g. a noun chunk, N_C, a finite verb chunk, FV_C, a prepositional chunk, P_C, etc.), its lexical head (identified by the label POTGOV) and any occurring modifier and preposition. It should be noted moreover that a chunked sentence does not contain information about the nature and the scope of inter-chunk dependencies. They are left to be parsed at further levels of analysis.

In the chunked sentence in Figure 1, use of underspecification is exemplified. The chunking process resorts to underspecified analyses in cases of systematic ambiguity, as adjectival modification. In fact, in Italian, post-nominal adjectives can be ambiguously interpreted either as restrictive modifiers or as secondary predicates. When a pre-nominal adjective such as *stesse* “the same” occur between the determiner (*le* “the”) and the noun (*disposizioni* “provisions”) it becomes part of a wider (nominal in this example) chunk. Post-nominal adjectives are instead regarded as independent chunks. It should be noted that in the case of the post-nominal reported in the example (*importato* “imported”) a second ambiguity type occurs, i.e. the one between adjective and past participle. This ambiguity is captured through the underspecified chunk category ADJPART_C, subsuming both an adjectival chunk and a participial chunk interpretation.

This underspecified approach to robust syntactic analysis of Italian texts has been proved to be fairly reliable. Lenci et al. (2001) provided a detailed evaluation of CHUG-IT parsing performance drawn on a corpus of financial newspapers articles. Results of automatic chunking were evaluated against a version of the same texts chunked by hand; they give a recall of 90.65% and a precision of 91.62%.

Starting from these results, in what follows we will provide an analysis of a corpus of Italian law texts by analysing the output of the chunking module included in *AnIta* CHUG-IT. *AnIta* is a general-purpose parsing system, which has already been tested as a component of the SALEM semantic annotation system of law texts with encouraging results (Bartolini et al., 2004b).

3.2. The corpora

For the construction of the Italian legislative corpora two different design criteria were taken into account, namely the regulated domain and the releasing agency. The corpus is made up of legal documents which a) regulate two different domains, i.e. the environmental and the consumer protection domains and b) which are released by three different agencies, i.e. European Union, national state and region.

3.2.1. The Environmental Corpus

The environmental corpus consists of 824 legislative, institutional and administrative acts for a total of 1,399,617 word tokens. It has been downloaded from the BGA (*Bollettino Giuridico Ambientale*), database edited by the Piedmont local authority for the environment[1]. The corpus includes acts released by three different agencies, i.e. European Union, Italian state and Piedmont region, which cover a nine-year period (from 1997 to 2005). It is a heterogeneous document collection (henceforth referred to as Environmental Corpus) 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.

3.2.1. The Consumer Law Corpus

The corpus containing legal texts which regulate the consumer protection domain is a more homogeneous collection. It is made up of 18 European Union Directives in consumer law (henceforth referred to as Consumer Law Corpus), for a total of 74,210 word tokens. Unlikely the Environmental Corpus, it includes only Italian European law texts.

3.3. Morpho-lexical analysis

As long as the accuracy of *chunking* depends on the accuracy of previous morphological analysis taken in input, text processing of legal sublanguage requires a general-purpose morphological lexicon tuned to domain-specific requirements. That includes updated information concerning grammatical categories (e.g. noun, verb, adjective, etc.), as well as morphological features (e.g. number, gender, person, etc.) for unknown words.

We started this update process from *AnIta*’s failures in the legal corpora analysis. Actually, we inspected the results obtained by the AnItamodule in charge of morpholog ical analysis, MAGIC (Battista and Pirrelli, 1999). MAGIC includes a general-purpose lexicon of about 100,000 lemmas. Thus, it has been necessary to “feed” it with domain-specific terms. The output of MAGIC consists of the association of each word form with all its possible lemmas, together with the morpho-syntactic features describing it. To be more concrete, a sentence such as *Il presente decreto stabilisce le norme per la prevenzione ed il contenimento dell’inquinamento da rumore* ... “This decree establishes the rules for prevention and control of noise pollution …” has been morphologically analysed as reported in Figure 2.

```
Il IL#RD@MS# IL#SP@NN#
presente   PRESENTIRE#V@S3IP#   PRESENTE#A@FS@MS#
PRESENTE#S@FS@MS#
decreto DECRETARE#V@S1IP# DECRETO#S@MS#
stabilisce STABILIRE#V@S3IP#
le LO#RD@FP# LE#PQ@FP3@FS3#
norme NORMA#S@FP#
per PER#E@#
la LO#RD@FS# LA#PQ@FS3# LA#S@MP@MS#
prevenzione PREVENZIONE#S@FS#
ed E#CC@#
il IL#RD@MS#
contenimento CONTENIMENTO#S@MS#
dell' DI#E@FS@MS#
inquinamento INQUINAMENTO#S@MS#
da DA#E@#
rumore RUMORE#S@MS#
```

Figure 2: MAGIC output

Interestingly, this preliminary phase of our study has shown that most words unknown to MAGIC lexicon belong to the lexicon of regulated domain, rather than to legal lexicon itself. In particular, 53.93% of the total number of MAGIC failures in the Environmental Corpus processing were due to chemical and physical lexicon particular to this special domain. On the contrary, processing the Consumer Law Corpus an insignificant percentage of failures resulted to be due to the consumer protection domain. As an application, a terminology extraction system in the legal field can benefit from a

[1] http://extranet.regione.piemonte.it/ambiente/bga/

linguistic study of the morpho-lexical features of a given-regulated domain.

3.4. A comparative syntactic analysis

This section reports on the results obtained by inspecting the syntactic chunked output of CHUG-IT. As a first step, we focussed on

1) the distribution of chunk types in the legal corpora,
2) the depth of chains of prepositional chunks.

We intended to deal with those linguistic features of law texts which may "feed" higher-level semantic analysis of law texts. For this purpose, the percentage distribution of chunk types within the Italian Legislative Corpus (i.e. the Environmental and the Consumer Law Corpus) was compared with the analysis of an Italian reference corpus, the PAROLE corpus (Marinelli et al., 2003), made up of about 3 million words including texts of different types (newspapers, books, etc.).

Chunk types	Italian Legislative Corpus				PAROLE corpus	
	Environmental Corpus		Consumer Law Corpus			
	Count	%	Count	%	Count	%
Adj/Participial_C	38607	3.56	1689	2.74	29218	1.90
Adjectival_C	126267	11.66	6146	10.00	65740	4.27
Adverbial_C	13021	1.20	1006	1.63	49038	3.19
Coordinating_C	59585	5.50	3095	5.03	73073	4.75
Finite Verbal_C	36838	3.40	3007	4.89	140604	9.14
Nominal_C	226529	20.92	13062	21.25	413821	26.92
Non finite verbal_C	19569	1.80	5867	9.54	41674	2.71
Predicative_C	13047	1.20	843	1.37	21772	1.41
Prepositional_C	321167	29.66	14152	23.03	338037	21.99
Punctation_C	192419	17.77	9756	15.87	278897	18.14
Subordinating_C	22026	2.03	2288	3.72	70226	4.56
Unknown_C	13439	1.24	535	0.87	14964	0.97

Table 1: Comparative distribution of chunk types.

3.4.1. Distribution of chunk types

In what follows, we will focus on the Italian legal language peculiarities which are worthy of being discussed. Chi-squared test applied on Table 1 and 2 confirms the existence of a significant correlation between corpus variation and chunk type distribution.

3.4.1.1. Prepositional and Nominal Chunks

Looking at Table 1 it can be noticed that prepositional chunks (Prepositional_C) are the most frequent chunk type within the whole Italian Legislative Corpus. On the contrary, nominal chunks (Nominal_C) are the most recurring chunk type within the reference corpus. However, it should be appreciated that prepositional as well as nominal chunks are differently distributed between the Environmental Corpus and the Consumer Law Corpus. Namely, in the Environmental Corpus prepositional chunks constitute 29.66% while the nominal chunks are 20.92%; in the Consumer Law Corpus the former ones are 23.03% while the latter ones are the 21.25%.

In-depth results have been figured out by carrying out further analysis of different distributions of prepositional and nominal chunks between legislative and ordinary language corpora. Table 2 shows that the heterogeneous composition of the Environmental Corpus affects the distribution of chunk types. Namely, the "regional" as well as the "state" Environmental sub-corpora show the highest occurrence of prepositional chunks (i.e. respectively 29.46% and 30.87%) and, on the contrary, the lowest frequency of nominal chunks (i.e. 21.17% and 20.18%).

Interestingly, it seems that the Italian European legal language has linguistic features which are more similar to those of ordinary language ones than the "regional" and "state" legal language. Table 2 shows in particular that prepositional chunks have lower occurrences, with respect to "regional" and "state" cases, both in the European Environmental sub-corpus (27.61%) and in the European directives making up the Consumer Law Corpus (23.03%). Nominal chunks as well have a higher frequency in both the European corpora (respectively, 22.10% in the European Environmental sub-corpus and 21.25% in the Consumer Law Corpus). Still, it is worth noting that the different regulated domains may have affected legal language, i.e. linguistic peculiarities specific to a given-regulated domain may have been assimilated. Nevertheless, the Italian European legal corpora still differ from the Italian reference corpus.

Chunk Types	Italian Legislative Corpus								PAROLE Corpus	
	Environmental Corpus						Consumer Law Corpus			
	Region		State		Europe					
	Count	%	Count	%	Count	%	Count	%	Count	%
Adj/Participial_C	7247	3.58	20305	3.58	11055	3.52	1689	2.74	29218	1.90
Adjectival_C	24949	12.33	68931	12.16	32387	10.33	6146	10.00	65740	4.27
Adverbial_C	2149	1.06	5944	1.04	4928	1.57	1006	1.63	49038	3.19
Coordinating_C	10315	5.09	31930	5.63	17340	5.53	3095	5.03	73073	4.75
Finite Verbal_C	5857	2.89	16601	2.92	14380	4.58	3007	4.89	140604	9.14
Nominal_C	42850	21.17	114404	20.18	69275	22.10	13062	21.25	413821	26.92
Non finite verbal_C	3509	1.73	7927	1.39	8133	2.59	5867	9.54	41674	2.71
Predicative_C	1850	0.91	6467	1.14	4730	1.50	843	1.37	21772	1.41
Prepositional_C	59615	29.46	175011	30.87	86541	27.61	14152	23.03	338037	21.99
Punctation_C	36373	17.97	103696	18.29	52350	16.70	9756	15.87	278897	18.14
Subordinating_C	3348	1.65	10068	1.77	8610	2.74	2288	3.72	70226	4.56
Unknown_C	4279	2.11	5496	0.96	3664	1.16	535	0.87	14964	0.97

Table 2: Comparative distribution of chunk types within the Italian Legislative Corpus accounting for three different agencies.

3.4.1.2. Verbal Chunks

The fairly low percentage of verbal chunks resulted to be one of the main specific features of law texts. When the Italian reference corpus has 9.14% of the finite verbal chunks, they are about a third within the "regional" and "state" law texts (2.89% and 2.92% respectively); the Italian European law texts show a halved percentage of occurrences (i.e. the finite verbal chunks are 4.58% in the Environmental European texts and 4.89% in the Consumer Law Corpus).

Interestingly, the reference corpus and the Environmental Corpus were less different concerning the occurrences of non-finite verbal chunks. On the contrary, the European Consumer Law Corpus has quite a high frequency of non-finite verbal chunks: they are 9.54%, with respect to 2.71% of the reference corpus.

This low presence of verbs within legal texts is an important issue not only from a linguistic point of view. Since verbs have a central role as connecting elements between concepts, in principle their low frequency may negatively affect the performance of an Event Extraction component for the semantic annotation of law texts or for the ontology learning process. If NLP-based, this higher-level layer should take into account the fact that few legal facts or events are linguistically realised through verbs. Rather they can be nouns, noun groups or other nominal constructions, typically embedded in PP-attachment chains. It could be the case that, for example, the event *accertare* "to verify" is realised in

a) a verbal construction, through the verb *accertare* "to verify"; e.g. *l'autorità amministrativa competente* ***accerta*** *la compatibilità paesaggistica* "the relevant administrative authority verifies the landscape compatibility";
b) a nominal construction, through the noun *accertamento* "verification"; in this case the event is linguistically realised as a nominal construction made up of PP-attachments; e.g. *l'autorità preposta alla gestione del vincolo ai fini dell'****accertamento*** *della compatibilità paesaggistica ...* "the authority in charge of the management of the obligation to the verification of the landscape compatibility ".

Exploring the semantic frames of legal events, the morpho-syntactic pre-processing of law texts may feed a layered analysis with "linguistic intelligence".

3.4.2. PP-attachment chains

In section 4.4.1.1., a higher occurrence of prepositional chunks within law texts was reported with respect to ordinary language. This may be due to deep PP-attachment chains which include a high number of embedded prepositional chunks.

In order to test this hypothesis, we have automatically computed the occurring PP-chains. We considered as a PP-chain the following typology of cases:

a) chains of consecutive prepositional chunks, such as *presentazione delle domande di contributo ai Comuni per l'attivazione dei distributori per la vendita di metano* "submission of contribution requests to Municipalities for the activation of distributors for the sale of natural gas", which is chunked as follows [$_{\text{N\_C}}$ presentazione] [$_{\text{P\_C}}$ delle domande] [$_{\text{P\_C}}$ di contributo] [$_{\text{P\_C}}$ ai Comuni] [$_{\text{P\_C}}$ per l'attivazione] [$_{\text{P\_C}}$ di distributori] [$_{\text{P\_C}}$ per la vendita] [$_{\text{P\_C}}$ di metano] (see Figure 3);
b) sequences of prepositional chunks with possibly embedded adjectival chunks, such as *disciplina del canone regionale per l'uso di acqua pubblica* "regulation of the regional fee for public water usage", which is chunked as follows [$_{\text{N\_C}}$ disciplina] [$_{\text{P\_C}}$ del canone] [$_{\text{ADJ\_C}}$ regionale] [$_{\text{P\_C}}$ per l'uso] [$_{\text{P\_C}}$ di acqua] [$_{\text{ADJ\_C}}$ pubblica];
c) sequences of prepositional chunks with possibly embedded adjectival chunks, coordinative conjunctions and/or "light" punctuation marks (i.e. comma), such as *acqua*

destinata all'uso igienico e potabile, all'innaffiamento degli orti ... "water devoted to sanitary and drinkable usage, to garden watering", which is chunked as follows [$_{\text{N\_C}}$ acqua] [$_{\text{ADJPART\_C}}$ destinata] [$_{\text{P\_C}}$ all'uso] [$_{\text{ADJ\_C}}$ igienico] [$_{\text{COORD\_C}}$ e] [$_{\text{ADJ\_C}}$ potabile] [$_{\text{PUNC\_C}}$,] [$_{\text{P\_C}}$ all'innaffiamento] [$_{\text{P\_C}}$ degli orti].

Following our comparative approach, we also focused on the length of PP-chains amongst different text types. We intended to compute how many embedded PP-attachments could occur within a sentence of law texts with respect to an ordinary language sentence. Table 3 shows the different distribution of PP-chains in the different kinds of corpora. Law texts appeared to have a higher percentage of deep PP-chains with respect to the reference corpus. It should be noticed that chains including from 5 to 11 embedded chunks are mainly involvd. For example, chains of 8 PP-attachments are 5.78% of the total amount of PP-chains occurring within the "regional" texts and 5.52% in the "state" texts. Yet, they are only 2.47% in the ordinary language texts. As already observed, the Italian European law texts have a midway behaviour between legal language and the ordinary one. It follows that within ordinary language texts deep PP-attachment chains do occur as well, but with a much lower frequency indeed.

"IL DIRIGENTE vista la l.r. 8_agosto_1997, n._51; … determina di riaprire, fissandoli al 29_ottobre_2004 , per le motivazioni di cui in premessa, **i termini per la presentazione delle domande di contributo ai Comuni per l' attivazione di distributori per la vendita di metano per autotrazione di cui al bando** approvato con D.d . n._505/22.4 del 26_novembre_2002 …"

```
[[CC:N_C][DET:IL#RD@MP][POTGOV:TERMINE#S@MP]]
[[CC:P_C][PREP:PER#E][DET:LO#RD@FS][POTGOV:PRESENTAZI
ONE#S@FS]]
[[CC:di_C][DET:LO#RD@FP][POTGOV:DOMANDA#S@FP]]
[[CC:P_C][PREP:DI#E][POTGOV:CONTRIBUTO#S@MS]]
[[CC:P_C][PREP:A#E][DET:IL#RD@MP][POTGOV:COMUNE#S@MP]
]
[[CC:P_C][PREP:PER#E][DET:LO#RD@FS][POTGOV:ATTIVAZION
E#S@FS]]
[[CC:P_C][PREP:DI#E][POTGOV:DISTRIBUTORE#S@MP]]
[[CC:P_C][PREP:PER#E][DET:LO#RD@FS][POTGOV:VENDITA#S@
FS]]
[[CC:P_C][PREP:DI#E][POTGOV:METANO#S@MS]]
[[CC:P_C][PREP:PER#E][POTGOV:AUTOTRAZIONE#S@FS]]
[[CC:P_C][PREP:DI#E][POTGOV:CUI#P@FP@FS@MP@MS]]
[[CC:P_C][PREP:A#E][DET:IL#RD@MS][POTGOV:BANDO#S@MS]]
```

Figure 3: A PP-chain of 11 consecutive prepositional chunks

The different distribution of PP-attachment chains in law texts and in the reference corpus appears to be mainly due to:

a) chains of embedded prepositional complements, such as the one reported in Figure 3;
b) chains of embedded cross-references to other texts or to single parts of them.

This elaborate syntactic structure may affect the whole law text comprehension. That is in line with some findings in studies on linguistic complexity, mainly in the cognitive and psycholinguistic field (see Fiorentino, 2007 for a survey of the state-of-the-art). It was figured out that our *short term memory* is able to receive, process and remember an average of 7 linguistic units. In processing a given input sentence the language user attempts to obtain closure on the linguistic units contained in it as early as possible. Thus, it is perceptually "costly" to carry on analysing deep chains of embedded sentence constituents. Syntactic complexities responsible for difficult processing of natural language sentences have also been studied.

In law texts, in particular, this is typically the case of the intra-textual reference and inter-textual cross-reference. Figure 4 shows how a cross-reference is typically realised within law texts. The bold part of the sentence reported in the excerpt below has been parsed as a deep PP-chain. The reference to other law texts sections as well as to other texts is syntactically realised through a deep chain of embedded chunks (mainly prepositional and adjectival).

"Dalla data di entrata in vigore del presente decreto agli impianti di incenerimento di cui all'articolo 1 non si applicano **le prescrizioni di cui al paragrafo 3.3 della deliberazione 27 luglio 1984 del Comitato interministeriale di cui all'articolo 5 del decreto del Presidente della Repubblica** 10 settembre 1982, n. 915, …"

```
[[CC:N_C][DET:LO#RD@FP][POTGOV:PRESCRIZIONE#S@FP]]
[[CC:P_C][PREP:DI#E][POTGOV:CUI#P@FP@FS@MP@MS]]
[[CC:P_C][PREP:A#E][DET:IL#RD@MS][POTGOV:PARAGRAFO#S@
MS]]
[[CC:ADJ_C][POTGOV:3.3#N]]
[[CC:di_C][DET:LO#RD@FS][POTGOV:DELIBERAZIONE#S@FS]]
[[CC:ADJ_C][POTGOV:27_luglio_1984#N]]
[[CC:di_C][DET:IL#RD@MS][POTGOV:COMITATO#S@MS]]
[[CC:ADJ_C][POTGOV:INTERMINISTERIALE#A@FS@MS]]
[[CC:P_C][PREP:DI#E][POTGOV:CUI#P@FP@FS@MP@MS]]
[[CC:P_C][PREP:A#E][DET:LO#RD@MS][POTGOV:ARTICOLO#S@M
S]]
[[CC:ADJ_C][POTGOV:5#N]]
[[CC:di_C][DET:IL#RD@MS][POTGOV:DECRETO#S@MS]]
[[CC:di_C][DET:IL#RD@MS][POTGOV:PRESIDENTE#S@MS]]
[[CC:di_C][DET:LO#RD@FS][POTGOV:REPUBBLICA#S@FS]]
```

Figure 4: A typical cross-reference realisation

PP-chains depth	Italian Legislative Corpus								PAROLE Corpus	
	Environmental Corpus						Consumer Law Corpus			
	Region		State		Europe					
	Count	%	Count	%	Count	%	Count	%	Count	%
4	2822	38.48	8924	37.42	4164	43.19	611	45.32	10240	54.72
5	1723	23.71	5366	22.50	2258	23.42	356	26.40	4621	24.68
6	1043	14.35	3505	14.69	1380	14.31	139	10.31	1999	10.68
7	612	8.42	2103	8.81	725	7.52	104	7.75	910	4.85
8	420	5.78	1318	5.57	409	4.24	44	3.26	464	2.47
9	248	3.41	813	3.40	237	2.45	28	2.07	206	1.09
10	151	2.13	652	2.73	161	1.67	23	1.70	112	0.59
11	91	1.35	350	1.46	92	0.95	10	0.74	74	0.39
12	63	0.88	244	1.02	69	0.71	7	0.51	39	0.20
13	30	0.42	167	0.70	39	0.40	9	0.66	28	0.14
14	19	0.32	147	0.61	37	0.38	5	0.37	17	0.09
15	18	0.28	79	0.33	27	0.28	1	0.07	6	0.03
16	11	0.25	62	0.25	26	0.27	6	0.44	5	0.02
17	6	0.09	40	0.16	5	0.05	1	0.07	3	0.01
18	3	0.05	31	0.12	4	0.04	3	0.22	2	0.01
19	3	0.04	24	0.10	3	0.03	0	0.00	1	0.00
20	2	0.02	23	0.09	4	0.04	1	0.07	3	0.00

Table 3: Comparative distribution of PP-attachment chains.

4. The English legal texts analysis

4.1. An English legislative corpus

This section reports on the results of a first study carried out on a collection of 18 English European Union Directives in consumer law. The corpus is made up of the English version of the Italian corpus in consumer law. As a starting point of the English Directives linguistic analysis, we used the output of *GENIATagger*, a NLP component carrying out part-of-speech tagging and chunking (Tsuruoka et al., 2005).

As for the linguistic analysis of the Italian legal language, our survey aims at showing the main syntactic peculiarities of the English legal texts with respect to a general language corpus. Asa reference corpus we used a sub-corpus of the Wall Street Journal made up of 1,138,189 words.

4.2. A comparative syntactic analysis

In what follows, we show some preliminary results of the analysis of chunked texts. In this case, we focused on the distribution of chunk types. A more exhaustive syntactic analysis is still ongoing, also including the analysis of PP-chains.

4.2.1. Distribution of chunk types

4.2.1.1. Prepositional and Nominal Chunks

The comparative distribution of chunk types between the English Legislative Corpus in consumer law and the reference corpus shows some already detected legal language peculiarities. As in the Italian legal texts, within the English legal documents the occurrence of prepositional chunks is higher than in the general language texts (see Table 4). Namely, they are 27.21% of the total chunk types in the English European Union Directives, while they are 19. 88% in the Wall Street Journal sub-corpus. At the same time, the percentage of nominal chunks is lower in legal texts (48.16%) than in the reference corpus, where they represent 51.84% of the identified chunks.

Chunk Types	English Legislative Corpus		WSJ Corpus	
	Count	%	Count	%
Nominal_C	17731	48.16	336635	51.84
Prepositional_C	10019	27.21	129131	19.88
Finite verbal_C	3378	9.17	101092	15.56
Non finite verbal_C	2401	6.52	26673	4.10
Adverbial_C	835	2.26	24139	3.71
Adjectival_C	823	2.23	11726	1.80

Table 4: Comparative distribution of chunk types.

4.2.1.2. Verbal Chunks

Regarding the distribution of verbal chunks, our survey has shown that they have quite a low percentage of occurrence. In particular, the finite verbal chunks represent 9.17% of the total chunk types within the English legislative corpus; while they are 15.56% in the reference corpus. In line with what it has been observed for the Italian European texts, the occurrence of non-finite verbal chunks is different; they are more frequent in legal texts (6.52%) than in the reference corpus (4.10%).

5. Comparing Italian and English legal language peculiarities

In order to figure out whether some syntactic peculiarities were shared by Italian and English legal language, we carried out a contrastive analysis. We compared the *GENIATagger* output with the CHUG-IT one. Namely, we analysed the chunked texts resulting by parsing the English and Italian European legal corpus on the consumer protection domain. It goes without saying that the outputs of the two different tools cannot be directly compared. Because of different grammatical requirements in the two languages considered (i.e. Italian and English), as well as because of different NLP techniques exploited, in some cases there are not exactly equivalent outputs and/or linguistic structures to be compared.

In particular, it should be noticed that the output of the *GENIATagger* and that of CHUG-IT mostly differ becuase of their representation of nominal and prepositional chunks. The fragment of *GENIATagger* chunked text, reported in Figure 5 below, shows how a nominal (i.e. NP) and a prepositional chunk (i.e. PP) have been parsed. A nominal chunk can be a textual unit of adjacent word tokens, such as *certain exonerating circumstances*, which includes an adjective (*certain*, JJ) at the beginning (B-NP), an introducing present participle (*exonerating*, VBG) and a common noun (*circumstances*, NNS) as two inner elements (I-NP). Yet, it can also be made up of a single word token, such as *proof*, which includes a common noun (NN) only, or *he*, which is made up of a personal pronoun (PRP). Moreover, a prepositional chunk does not contain anything more than the introducing preposition itself, such as *to* or *as* (which is also labelled as preposition other than *to* and subordinating conjunction, i.e. IN). That is, the governing scope of a preposition never goes beyond the limits of the prepositional chunk it appears in. This processing strategy is relevant for the English syntactic features concerning the stranding of prepositions within the sentence. It follows that a noun is excluded from the list of lexical heads within prepositional chunks: the governing element of the chunk is always a preposition. On the contrary, in CHUG-IT prepositions with the first adjacent element. As section 4.1. reports, a noun thus can be the lexical head (POTGOV) of a prepositional chunk; while preposition is one of the inter-chunk word tokens embedded in (i.e. *ad un prodotto*, "to a product", is parsed as `[[CC:P_C] [`**`PREP:AD#E`**`] [DET:UN#RI@MS] [POTGOV:PRODOTTO#S@MS]])`. Thuș prepositions can never be parsed alone since they are part of a distinct chunk.

"…if he furnishes proof as to the existence of certain exonerating circumstances …"

```
if   if   IN   B-SBAR   O
he   he   PRP  B-NP     O
furnishes     furnish  VBZ  B-VP     O
proof    proof    NN   B-NP     O
as   as   IN   B-PP     O
to   to   TO   B-PP     O
the  the  DT   B-NP     O
existence     existence     NN   I-NP     O
of   of   IN   B-PP     O
certain  certain  JJ  B-NP     O
exonerating  exonerate    VBG  I-NP     O
circumstances circumstance NNS  I-NP     O
```

Figure 5: *GENIATagger* output

5.1. Two shared syntactic features

Despite these different chunked representations, the maindistribution of chunk types between Italian and English European legal texts shows similarities. Comparing the two European legal languages, some features have been revealed to be shared. Table 5, shows that they mainly are:

- an high occurrence of prepositional chunks,
- a fairly low presence of finite verbal chunks.

It should be appreciated that the prominence of this Italian-English contrastive analysis is given by the joint comparison with a general language reference corpus. The results of the linguistic analyses carried out on European legal texts in both languages should not be read as absolute values; rather they should be interpreted with respect to the results obtained on each corresponding open-domain corpus.

Chunk Types	Italian Legislative Corpus				PAROLE Corpus		English Legislative Corpus		WSJ Corpus	
	Environmental Corpus		Consumer Law Corpus							
	Count	%	Count	%	Count	%	Count	%	Count	%
Nominal_C	226529	20.92	13062	21.25	413821	26.92	17731	48.16	336635	51.84
Prepositional_C	321167	29.66	14152	23.03	338037	24.86	10019	27.21	129131	19.88
Finite verbal_C	36838	3.40	3007	4.89	140604	9.14	3378	9.17	101092	15.56
Non finite verbal_C	19569	1.80	5867	9.54	41674	2.71	2401	6.52	26673	4.10

Table 5: Distribution of chunk types amongst Italian and English law texts

Concerning the verbal behaviour of legal texts, it is worth noticing an English European legal language specificity. The English European Union Directives have shown a higher percentage of modal auxiliaries (e.g. *may*, *can*, *could*, etc.) with respect to the Wall Street Journal sub-corpus. Table 6 below shows the different percentage occurrences amongst the two different languages. It should be noticed that the percentage of modal auxiliaries is higher throughout Italian legal language (i.e. within both the Environmental Corpus and the Consumer Law Corpus) with respect to ordinary Italian language. Nevertheless, modal auxiliaries occur far more often within English European legal language with respect both to Italian European legal language and ordinary English language.

	Legal language				Ordinary language	
Italian	Environmental Corpus		Consumer Law Corpus		Count	%
	Count	%	Count	%	11237	5.50
	7269	11.57	702	12.48		
English	Count		%		Count	%
	1324		22.59		12924	9.99

Table 6: Occurrences of modal auxiliaries.

The accurate recognition of modal information (i.e. modality) is vital for a correct interpretation of sentences as well as of facts, e.g. a speculative statement should be taken very differently from one expressing a definite fact. Even though the use of speculation is particularly common within the experimental life sciences (Medlock and Briscoe, 2007), a fine-grained content analysis of law texts could benefit from a correct recognition of modality, mainly deontic. As an application, a legal facts extraction systems endowed with "linguistic intelligence" should be capable of recognising different linguistic realisations of modality information.

6. Conclusion and future work

We have presented an analysis of some legal language syntactic features detected within Italian legal corpora. For this purpose we used NLP techniques, mainly exploiting a shallow parsing approach, i.e. chunking. Not "balking" at domain-specific constructions as well as at particularly complex syntactic constructions that do not follow general grammar rules, chunking can be a reliable starting point for parsing texts as complex as law texts.

We carried out a comparative study in order to investigate to what extent legal language differs from ordinary language. The occurrence of prepositional chunks as one of the more visible syntactic phenomena has been mainly considered. It has been observed that that it may be due to the presence of deep PP-attachment chains which include a high number of embedded prepositional chunks. Finally, it has been found that "regional" and "state" law texts differ from ordinary language more than Italian European law texts.

Moreover, a contrastive approach has also been followed. We aimed at finding out whether some of the syntactic peculiarities of law texts previously detected were shared by English law texts. For this purpose, we exploited a different battery of NLP tools, still relying on the same syntactic chunking technique. We first investigated whether and to what extent English legal language differs from ordinary English language; secondly we compared the obtained results with the Italian case.

We can think about a quite great potential for this linguistic analysis of legal texts. A very promising topic we are going to address is that of using the information provided by typical chunk patterns to improve a terminology extraction task. CHUG-IT results showed that Italian multi-word terms are typically realised as syntactic structures made up of several adjacent prepositional chunks. A strategy of complex terms extraction from law texts could be refined exploiting information tuned on domain-specific linguistic peculiarities, i.e. improving the acquisition process coverage.

Far from being simply a stylistic remark, an investigation through the distribution of deep PP-chains could improve further semantic analyses. Bartolini et al., (2004b) in their study on semantic annotation of legal texts suggested that, for example, the semantic mark-up of *modifications* and *obligations* should require linguistic knowledge of the syntactic structures underlying the Italian provisions text. Our current study has shown that legal entities (i.e. logical components or actors) involved are linguistically represented through deep PP-attachment chains. It is still an ongoing study aimed at verifying whether a robust syntactic pre-processing of textual *modifications* can improve an Italian legal information extraction system. Knowing how, for example, the entities involved in the text of amendment (*novella*) and the amending text (*novellato*) are linguistically realised (i.e. the syntactic patterns which they represented in) can improve their semantic mark-up.

Finally, our further directions of research are devoted to taking into account the low occurrences of verbs (mainly finite) detected in law texts. This makes challenging the annotation as well as the extraction of semantic frames representing legal events in text. The linguistic analysis carried out showed that not only events described by verbs, but also by nominalised verbs and possibly by other nominal constructions should be considered. In this study, we intended to suggest that an event extraction system could benefit by the inspection of linguistic structures representing events and entities which take part.

7. Acknowledgements

This work is the result of research activities of the Dylan Lab (Dynamics of Language Laboratory) of the Istituto di Linguistica Computazione (ILC-CNR) of Pisa and of the Department of Linguistics (Computational Linguistics division) of the University of Pisa. *AnIta* has been developed and provided by its members.

8. References

Abney S.P., Parsing by chunks, in R.C. Berwick et al (eds.), Principled-Based Parsing: Computation and Psycholinguistics, Kluwer Academic Publishers, Dordrecht, 1991, pp. 257-278.

Bartolini R., Lenci A., Montemagni S., Pirrelli V., Hybrid Constraints for Robust Parsing: First Experiments and

Evaluation, in Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation, 26-28 May 2004, Centro Cultural de Belem, Lisbon, Portugal, pp. 795-798.

Bartolini R., Lenci A., Montemagni S., Pirrelli V., Soria C., Automatic classification and analysis of provisions in legal texts: a case study. In R. Meersman, Z. Tari, A. Corsaro (eds.) On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops. Lecture Notes in Computer Science, 3292/2004, Springer-Verlag. 593-604.

Battista M., Pirrelli V. 1999, Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane, *ILC-CNR*.

Buitelaar P., Cimiano P. and Magnini B., Ontology Learning from text: an Overview, in Buitelaar et al., (eds.), Ontology Learning from Text: Methods, Evaluation and Applications (Volume 123 Frontiers in Artificial Intelligence and Applications): 3-12, 2005.

Federici S., Montemagni S., Pirrelli V., 1996 Shallow Parsing and Text Chunking: a View on Underspecification in Syntax in J. Carroll (Ed.), *Proceedings of the Workshop On Robust Parsing*. 12-16 August 1996, ESSLLI, Prague.

Fiorentino G., (2007) Web usability e semplificazione linguistica. In: F. Venier (a cura di), *Rete Pubblica. Il dialogo tra Pubblica Amministrazione e cittadino: linguaggi e architettura dell'informazione*, Perugia, Edizione Guerra: 11-38.

Lame G., Using NLP techniques to identify legal ontology components: concepts and relations, Lecture Notes in Computer Science, volume 3369: 169-184, 2005.

Lenci A., Montemagni S., Pirrelli V. 2001, CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation, in Linguistica Computazionale, Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma.

Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A., The Italian PAROLE corpus: an overview, in A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVI-XVII, (2003). Pisa-Roma, IEPI. Tomo I, 401-421.

Moens M-F., Mochales Palau R., Boiy E., Reed C., Automatic detection of Arguments in Legal Texts, in Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL), June 4-8, 2007, Stanford Law School, Stanford, California.

Reeve, L., and Han, H. (2005). Survey of Semantic Annotation Platforms. Proceedings of the 20th Annual ACM Symposium on Applied Computing, Web Technologies and Applications track, Santa Fe, New Mexico.

Saias J. and Quaresma P., A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System, Lecture Notes in Computer Science, Volume 3369, 2005, pp. 185-200.

Tsuruoka Y., Tateishi Y., Kim J-D., Ohta T., McNaught J., Ananiadou S. and Tsujii J., Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005

Van Gog, R. and Van Engers, T.M., Modelling Legislation Using Natural Language Processing, in Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics (SMC2001).

Walter S. and Pinkal M., Automatic extraction of definitions from german court decisions, in Proceedings of the COLING-2006 Workshop on "Information Extraction Beyond The Document", Sydney, July 2006, pp. 20-28.

Automatic Identification of Legal Terms in Czech Law Texts

Karel Pala, Pavel Rychlý, Pavel Šmerk

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{pala,pary,xsmerk}@fi.muni.cz

Abstract

Law texts including constitution, acts, public notices and court judgements form a huge database of texts. As many texts from small domains, the used sublanguage is partially restricted and also different from general language (Czech). As a starting collection of data the legal database Lexis containing approx. 50,000 Czech law documents has been chosen. Our attention is concentrated mostly on noun groups which are the main candidates for law terms. We were able to recognize 3992 such different noun groups in the selected text samples. The paper also presents results of the morphological analysis, lemmatization, tagging, disambiguation, and the basic syntactic analysis of Czech law texts as these tasks are crucial for any further sophisticated nanural language processing. The verbs in legal texts have been explored preliminarily as well. In this respect we are trying to explore how the linguistic analysis can help in identification of the semantic nature of law terms.

1. Introduction

In the paper we describe the first results of the new project whose final goal is to build an electronic dictionary of Czech law terms. We have started with a legal database Lexis developed at the Institute of Law, Czech Academy of Sciences, which presently includes approx. 50,000 Czech law documents ranging from the beginning of Czechoslovak Republic in 1918 to present days. It also includes court judgements, main representative law textbooks and law reports. All the texts exist in electronic form.

The first part of the paper presents results of the preparation step for the subsequent term identification – the morphological analysis. For this purpose we have used the tools developed in the Natural Language Processing Centre of the Faculty of Informatics, Masaryk University, particularly, morphological analyser Ajka (Sedláček, 2005) performing lemmatization and tagging and a new tool for grammatical disambiguation named DESAMB (Šmerk, 2007). The tools have been designed for general Czech but it appears that they can be exploited for law sublanguage with some minor modifications, namely adding law terms. The tools are now configured to analyze all Czech law texts contained in the Lexis database, the presented results from the pilot project are described below.

In the second part we report about term identification via syntactic analysis which has used tool DIS/VADIS (Žáčková, 2002), a partial parser for Czech. As a result list of noun groups has been obtained that can be considered as good candidates for law terms. We are also having a look at the verbs existing in law texts because they are relational elements linking together the established law terms. Here the apparatus of valency frames (Horák and Hlaváčková, 2005) comes as an appropriate instrument. It allows us to explore context patterns in which law terms occur and see how they behave in the law text.

The general goal is to find out in what extent linguistic analysis can contribute to semantic analysis of the law text. We are at the beginning of this enterprise.

1.1. Pilot project

As a pilot project we have decided to analyse the current version of the Penal Code of the Czech Republic. It is one of the biggest law documents containing almost 36,000 word forms. The overall characteristic of the document can be found in Table 1.

Number of	
word forms (tokens)	35,893
numbers	2,647
punctuation marks	9,135
tokens total	47,865
different word forms (types)	5,019
different numbers	467
different punctuation marks	12
types total	5,019

Table 1: The overall characteristic of the Penal Code of the Czech Republic

The task is to process the document by the Czech morphological analyser (lemmatizer) Ajka in such a way that for each word form in the source text a morphological information in the form of morphological tags is obtained. Thus we get information to what parts of speech the word forms belong, and, for instance, for nouns also grammatical categories like gender, number and case. Each word form in the document is associated with its respective lemma as well. In the highly inflectional language like Czech all this information is relevant for the further analysis of law terms. The results of the morphological analysis and lemmatization are transformed into a special format which is described below.

2. Morphological Analysis

We have used several simple scripts to create what is called vertical file from the source text. It is a plain text file without any formatting (word-processing options). Words are written in a column, i.e. each line contains one word, number or punctuation mark. Optional annotation is on the same line and the respective words are divided by the tabulator character. The first step uses only word forms from the source text. The vertical file serves as an input text for many corpus processing tools like CQP (Schulze and Christ, 1996) and Manatee (Rychlý, 2000).

In the next step, we processed the vertical file with the morphological analyser Ajka (Sedláček, 2005). It is a tool exploited for annotating and lemmatizing general Czech texts, however, the processing law texts requires modifications, e.g. enriching the list of stems of Ajka. The programme yields all possible combinations of lemma and morphological tags for each Czech word form.

Table 2 presents an example of the Ajka output, the tag **k1gFnSc1** means: part of speech (**k**) = noun (**1**), gender (**g**) = female (**F**), number (**n**) = singular (**S**) and case (**c**) = first (nominative) (**1**), tags beginning with **k2** are adjectives, **k3** – pronouns, **k5** – verbs and **k7** – prepositions.

As one can see, many word forms are ambiguous: there are more than one possible tag or even lemma for a given word form. In the analysed document, 76 % of word forms are ambiguous, more than 42 % of word forms have more than one possible lemma and average number of tags for an ambiguous word form is 6.75.

We have used part-of-speech tagger Desamb (Šmerk, 2007) to disambiguate such word forms. The output of the Desamb tool contains only the most probable lemma/tag for each word form. Table 3 contains output of Desamb for the input text above.

The annotated version of the document contains 2,560 different lemmas. Frequencies of each part of speech are in Table 4.

Příprava	příprava	k1gFnSc1
k	k	k7c3
trestnému	trestný	k2eAgInSc3d1
činu	čin	k1gInSc3
je	být	k5eAaImIp3nS
trestná	trestný	k2eAgFnSc1d1
podle	podle	k7c2
trestní	trestní	k2eAgFnSc2d1
sazby	sazba	k1gFnSc2
stanovené	stanovený	k2eAgFnSc2d1
na	na	k7c4
trestný	trestný	k2eAgInSc4d1
čin	čin	k1gInSc4

Table 3: The document in vertical format with morphological annotation (after disambiguation)

Part of Speech	Count
k1 – noun	12884
k2 – adjective	4634
k3 – pronoun	2252
k4 – numeral	1028
k5 – verb	4504
k6 – adverb	933
k7 – preposition	3600
k8 – conjunction	3764

Table 4: Frequencies of part of speech in the document

3. Noun Groups

For the recognition of the noun groups we have used the partial syntactic analyzer for Czech DIS/VADIS (Žáčková, 2002) at first. Unfortunately, DIS/VADIS presently does not contain rules which can recognize genitival and coordinate structures because during the development of DIS/VADIS these rules were found too erroneous (overgenerating) when applied to an unrestricted text. However, there are plenty of such structures in the law texts and overgenerating is not a problem here because the results will be checked manually.

Moreover, the partial syntactic analyzer DIS/VADIS has one more disadvantage: it is written in Prolog which implies that the recognition process is rather slow. Therefore we have rewritten the rules for noun groups to Perl 5 regular expressions (which have nontrivial backtracking capabilities) and added the rules for genitival and coordinate structures and some adverbials common to the law texts which also were not recognized by DIS/VADIS (e.g. *zvlášť* (exceedingly), *zjevně* (evidently) etc.).

For each noun group found in the law texts we determine its:

Příprava	<l>příprava <c>k1gFnSc1 (preparation)
k	<l>k <c>k7c3 (to)
trestnému	<l>trestný <c>k2eAgMnSc3d1 <c>k2eAgInSc3d1 <c>k2eAgNnSc3d1 (criminal)
činu	<l>čin <c>k1gInSc3 <c>k1gInSc6 <c>k1gInSc2 <l>čina <c>k1gFnSc4 (act)
je	<l>být <c>k5eAaImIp3nSrDaI <l>on <c>k3p3gMnPc4xP <c>k3p3gInPc4xP <c>k3p3gNnSc4xP <c>k3p3gNnPc4xP <c>k3p3gFnPc4xP <l>je <c>k0 (is)
trestná	<l>trestný <c>k2eAgFnSc1d1 <c>k2eAgFnSc5d1 <c>k2eAgNnPc1d1 <c>k2eAgNnPc4d1 <c>k2eAgNnPc5d1 (criminal)

Table 2: Output of the morphological analyser Ajka

1. base form (nominative singular),
2. head
3. for nouns in genitive groups also their part.

For example for the noun group *dalším páchání trestné činnosti (subsequent commission of criminal activity, dative)* we get:

1. *další páchání trestné činnosti*
2. *páchání*
3. *další páchání*

We can recognize 8,594 noun groups counting repeating occurencies, 3,992 different noun groups. Table 5 lists several most frequent noun groups with the respective number of occurrences in the pilot data (since there are some conceptual problems with finding the correct English equivalent terms we do not offer them here).

The noun groups was analyzed and the respective 'base' of each noun group was derived. Due to the inflectional feature of Czech this cannot be done by simple lemmatization of all words in a noun group. The automatic transformation algorithm works in following steps:

- find dependences between parts (words of subgroups) of a noun group,
- locate the root – key word,
- identify matching noun group pattern,
- generate the correct word forms with matching gramatical categories.

The result of this algoritm are base forms of noun groups and they will appear as headwords in the final electronic dictionary. The most frequent base forms with respective number of occurrences in the pilot data are listed in Table 6.

Noun Group	Count
odnětím svobody	492
peněžitým trestem	139
jeden rok	123
trestný čin	79
odnětí svobody	76
účinnosti dne	65
zákazem činnosti	64
trestného činu	58
velkého rozsahu	49
závažný následek	47
zvlášť závažný následek	46
(jiné) majetkové hodnoty	46
těžkou újmu	44
značný prospěch	40
jiný zvlášť závažný následek	40
výjimečným trestem	39
organizované skupiny	39
člen organizované skupiny	39
značnou škodu	38

Table 5: The most frequent noun groups

Table 7 presents the most frequent part-of-speech patterns of the recognized noun groups. There are two counts in the table, 'Count Tokens' is the total number of occurrences of the respective pattern in the pilot data, 'Count Types' is the number of different noun groups matching such pattern.

4. Verb List

Though law terms typically consist of the nouns, noun groups and other nominal constructions we also have paid attention to the verbs found in the whole database of the 50,000 law documents. The reason for this comes from the fact that verbs on one hand do not display strictly terminological nature but on the other they are relational elements linking the terminological nouns and noun groups together. This can be captured by the surface and deep verb valency frames (Horák and Hlaváčková, 2005) of the verbs occuring in the law documents. We are not aware of any attempt to

Part of Speech Patterns	Count Tokens	Count Types
k2 – k1gI	1588	344
k2 – k1gF	1130	365
k1gN – k1gF	765	96
k2 – k1gN	478	213
k1gI – k1gN	204	57
k1gN – k1gI	203	80
k1gI – k1gF	195	67
k2 – k1gM	176	71
k2 – k2 – k1gF	163	65
k1gF – k1gI	162	48

Table 7: The most frequent POS patterns

Noun Group	Count
odnětí svobody	568
trestný čin	228
peněžitý trest	152
jeden rok	123
zákaz činnosti	81
trest odnětí svobody	69
účinnost dne	65
(jiná) majetková hodnota	65
velký rozsah	64
těžká újma	58
výjimečný trest	51
organizovaná skupina	49
závažný následek	47
zvlášť závažný následek	46
veřejný činitel	46
značný prospěch	40
jiný zvlášť závažný následek	40
značná škoda	39
člen organizované skupiny	39

Table 6: The most frequent terms

describe the valency frames of the verbs coming from law texts. Presently the verb list comprises 15,110 items, particularly 10,190 infinitives and 4,920 participles (which are mostly the passive ones). The list has been processed by the morphological analyzer Ajka (Sedláček, 2005) as a result we have obtained the list of 914 items that were not recognized by Ajka morphological tool. The structure of this list shows that at least three types of the non-recognized items can be observed:

1. erroneous forms caused by typing errors, they can be corrected, e. g. *cítit (feel),*

2. the verbs that Ajka does not know, i. e. the ones that do not appear in the Ajka's list of stems. Typically, they display a terminological character and they should be added to the Ajka's stem list, e. g. *derogovat (derogate)*. They will enrich the list of (Czech) stems and their law meanings constitute a terminological subset of verbs,

3. erroneous forms that cannot be corrected without correcting the whole paragraph of a law document (we do not touch them).

The next step is to add the non-recognized verbs to Ajka's list of the verb stems and to make an intersection with our existing database Verbalex (Horák and Hlaváčková, 2005) containing presently about 11,306 (general) Czech verbs.

5. Context patterns in law texts

We take the position that the decisive information about the semantics of the law terms comes from the contexts in which they occur. There are two ways how to approach this:

- To use statistical techniques by means of which we obtain the interesting contexts – they can be sorted and the semantic clusters they create can be built. The limitation here is that the data from the law texts are not large enough and in some cases we do not get enough contexts to make the necessary generalizations.

- To explore the valency frames in the law texts and find the semantic roles that are typical for the verbs in the law texts. We already have done this for approximately 11 000 of (general) Czech verbs and the result is that we learn enough not only about the verb meanings but also about arguments constituting the argument-predicate structure of the 'law' verbs.

We expect that the inventory of the semantic roles for 'law' verbs will reasonably differ from the 'general' verbs and, on the other hand, that there will be interesting polysemy which we capture by means of semantic roles occurring in the found valency frames of the 'law' verbs. The two approaches, obviously, can be combined.

To show how we understand valency verb frames and the corresponding semantic roles we offer the example with the two following verbs:

1. *uložit trest někomu (to condemn sb to a sentence)* The meaning of this verb can be described by the following frame:
   AG(judge:1)[1] – uložit – PAT(person:1)[3] ACT(sentence:1)[4]

2. *obvinit z trestného činu koho (accuse sb of criminal act)* The meaning of this verb can be captured by the following frame: AG(public prosecutor:1)[1] – obvinit – PAT(person:1)[4] ACT(act:2)[z2]

To explain briefly the notation used: for the semantic roles we use labels like AG(judge:1), which say that the agent of the verb *uložit (condemn)* has to be a judge, the second role can be any person and the third one is the ACT, i.e. a sentence. Moreover, the labels used for the roles are closely linked to Princeton WordNet literals and they represent nodes in this semantic network which yields relevantinformation about their senses. The numbers following the roles express morphological cases that have to be indicated in Czech.

The frames adduced here are very similar to the frames as they presently exist in our verb frame database Verbalex mentioned above. This means that the effort put into its building can be exploited also in the area of the law texts. The more important thing, however, is that the valency frames capture noun and prepositional groups obtained via morphological and syntactic analysis mentined above and tell us what is their meaning. In other words, this knowledge allows us to find out what entities are denoted by noun and prepositional groups in law text and on this ground to build an ontology for the law domain. Then it can be compared with the already existing law ontologies such as the one built within the LOIS (Lexical Ontologies for Legal Information Society) project[1]. In this project the ontology is built in the WordNet fashion. It can be expected that the ontology exploiting semantic roles in valency frames should be closer to law texts in their natural form. Thus we can conclude that building valency frames of verbs occuring in law text is one of the important tasks set in this project.

6. Conclusion

We have presented the preliminary results of the computational analysis of Czech law documents, or more precisely, their selected samples. On one hand we have used the already existing tools such as Ajka or DIS/VADIS, on the other hand we have modified them respectively for the purpose of the present task. As a result we can enrich them with regard to the law language but, more importantly, we have obtained basic knowledge about the grammatical structure of the law texts (law terminology) and in this way we are prepared to continue our exploration of the contexts in which law terms occur in the law documents.

The knowledge of such contexts is a necessary condition for a deeper understanding of how law terminology works and how it can be made more consistent. As an application we intend to obtain the basic rules for intelligent searching law documents. A tool based on such rules can serve to judges, attorneys and experts in creating new law documents. In other words, the relevant output of this work thus will be an electronic dictionary of law terms.

Acknowledgements

This work has been partly supported by the Academy of Sciences of the Czech Republic under the projects 407/07/0679 and by the Ministry of Education of the Czech Republic within the Centre of basic research LC536.

7. References

A. Horák and D. Hlaváčková. 2005. VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages, Third International Seminar*, pages 107–115, Bratislava. VEDA.

Pavel Rychlý. 2000. *Corpus managers and their effective implementation*. Phd thesis, Faculty of Informatics, Masaryk University.

B. M. Schulze and O. Christ, 1996. *The CQP User's Manual.*

Radek Sedláček. 2005. *Morphemic Analyser for Czech.* Ph.D. thesis, Masaryk University.

Pavel Šmerk. 2007. *Towards Morphological Disambiguation of Czech.* Ph.D. thesis proposals, Faculty of Informatics, Masaryk University.

Eva Žáčková. 2002. *Partial syntactic analysis of Czech.* Phd thesis, Faculty of Informatics, Masaryk University.

[1] see http://nlpweb.kaist.ac.kr/gwc/pdf2006/50.pdf and also http://www.ittig.cnr.it/Ricerca/materiali/lois/WhatIsLOIS.htm

Multilevel Legal Ontologies

**Gianmaria Ajani[1], Guido Boella[2], Leonardo Lesmo[2],
Alessandro Mazzei[2], Daniele Radicioni[2], Piercarlo Rossi[3]**

[1]Dipartimento di Scienze Giuridiche - Università di Torino,
[2]Dipartimento di Informatica - Università di Torino,
[3]Dipartimento di Studi per l'Impresa e il Territorio, Università del Piemonte Orientale
gianmaria.ajani@unito.it, {guido,lesmo,mazzei,radicion}@di.unito.it,
piercarlo.rossi@eco.unipmn.it

Abstract

In order to manage the knowledge representation of European law we have proposed the Legal Taxonomy Syllabus methodology (Ajani et al., 2007a; Ajani et al., 2007b). In this paper we apply such a methodology to a new issue, concerning the recasting of the Community legislation on contract law. In particular we show how to include in the knowledge representation system of the Legal Taxonomy Syllabus the Acquis Principles, which have been sketched by scholars in European Private Law from the so-called Acquis communautaire.

Index Terms: Lightweight Ontologies, European Directives, Acquis Principles.

1. Introduction

European Union Directives (EUDs) are sets of norms that have to be implemented by the national legislations and translated into each of the Member States' languages. The general problem of multilingualism in European legislation has recently been addressed by using linguistic and ontological tools, e.g. (Giguet and Luquet, 2006; Després and Szulman, February 2007; Casanovas et al., 2005; Vossen et al., 1999; Tiscornia, 2007). The management of EUD is particularly complex, since the implementation of a EUD does not correspond to a straight transposition into a national law. By converse, managing this kind of complexity with appropriate tools can facilitate the comparison and harmonization of national legislation (Boer et al., 2003).

For instance, LOIS Project aims at extending EuroWordnet with legal information, thus adopting a similar approach to multilinguism, with the aim at connecting legal ontology to a higher level ontology (Tiscornia, 2007). In previous works, we proposed the Legal Taxonomy Syllabus[1] (LTS), a tool to build multilingual conceptual dictionaries aimed at representing and analysing terminologies and concepts from EUDs (Ajani et al., 2007a; Ajani et al., 2007b).

Whilst the final goal of LOIS is to support applications concerning information extraction, the Syllabus is concerned with the access of human experts to the EU documents. LTS is based on the distinction between *terms* and *concepts*. The latter ones are arranged into ontologies that are organised in levels. In (Ajani et al., 2007b) only two levels were defined: the European level –containing only one ontology deriving from EUDs annotations–, and the national level –hosting the distinct ontologies deriving from the legislations of EU member states–. In this paper we add a further level into this schema, coming from the characterization of the *Acquis Principles*.

The Acquis Principles have been sketched by scholars in European Private Law from the so-called *Acquis communautaire*, the existing body of EU primary and secondary legislation as well as the European Court of Justice decisions (Ajani and Schulte-Nölke, 2003). Notwithstanding the importance of this existing body of settled laws, the Acquis is also a far wider notion, encompassing an impressive set of principles and obligations, going far beyond the internal market and including areas, such as agriculture, environment, energy and transports. In February 2003, the European Commission published an Action Plan aimed at achieving greater coherence in European contract law, by adopting a non-binding Common Frame of Reference (CFR) (European Commission, 2003). Accordingly, some areas of Acquis are currently in the way to be consolidated to enhance coherence in their implementation by the Member States, and their interpretation by courts.

Focus of the present paper is the integration of the Acquis principles as a further *level* of the LTS framework. The paper is structured as follows. In Section 2. we point out two main problems raised in comparative law as regards as EUDs and their transpositions. In Section 3. we describe how the methodology of the LTS allows to cope with these problems. In section 4. we describe the *Acquis level* and illustrate how the LTS can be enriched to account for the Acquis Principles (Research Group on the Existing EC Private Law, 2007), as well. Finally, in Section 5. we provide some conclusions and elaborate about future works.

2. Terminological and conceptual misalignment

Comparative Law has identified two key points in dealing with EUD, which make more difficult dealing with the polysemy of legal terms: we call them the *terminological* and *conceptual misalignments*.

In the case of EUD (usually adopted for harmonising the laws in the Member States), the terminological matter is complicated by the need to implement them in the national

[1] LTS is a dictionary of Consumer Law, which has been carried out within the broader scope of the Uniform Terminology Project, http://www.uniformterminology.unito.it (Rossi and Vogel, 2004). The current version of our system can be found at the address: www.eulawtaxonomy.org.

legislations. In order to have a precise transposition in a national law, a Directive may be subject to further interpretation. Thus, a *legal concept* can be expressed in different ways in a Directive and in its implementing national law. A single concept in a particular language can be expressed in a number of different ways in a EUD and in the national law implementing it. As a consequence we have a terminological misalignment. For example, the concept corresponding to the word *reasonably* in English, is translated into Italian as *ragionevolmente* in the EUD, and as *con ordinaria diligenza* into the transposition law.

In the EUD transposition laws a further problem arises from the different national *legal doctrines*. A legal concept expressed in an EUD may not be present in a national legal system. In this case we can talk about a conceptual misalignment. To make sense for the national lawyers' expectancies, the European legal terms have not only to be translated into a sound national terminology, but they also need to be correctly detected when their meanings are to refer to EU legal concepts or when their meanings are similar to concepts which are known in the Member states. Consequently, the transposition of European law in the parochial legal framework of each Member state can lead to a set of distinct national legal doctrines, that are all different from the European one. In case of consumer contracts (like those concluded by the means of distance communication as in Directive 97/7/EC, Art. 4.2), the notion to provide in a *clear and comprehensible manner* some elements of the contract by the professionals to the consumers represents a specification of the information duties which are a pivotal principle of EU law.

Despite the pairs of translation in the language versions of EU Directives (e.g., *klar und verständlich* in German - *clear and comprehensible* in English - *chiaro e comprensibile* in Italian), each legal term, when transposed in the national legal orders, is influenced by the conceptual filters of the lawyers' domestic legal thinking. So, *klar und verständlich* in the German system is considered by the German commentators referring to three different legal concepts: 1) the print or the writing of the information must be clear and legible (*gestaltung der information*), 2) the information must be intelligible by the consumer (*formulierung der information*), 3) the language of the information must be the national of consumer (*sprache der information*). In Italy, the judiciary tend to control more the formal features of the concepts 1 and 3, and less concept 2, while in England the main role has been played by the concept 2, though considered as plain style of language (not legal technical jargon) thanks to the historical influences of plain English movement in that country. Note that this kind of problems identified in comparative law has a direct correspondence in the ontology theory. In particular Klein (Klein, 2001) has remarked that two particular forms of ontology mismatch, are *terminological* and *conceptualization* ontological mismatch which straightforwardly correspond to our definitions of misalignments.

In next Section we describe how the LTScan help to properly manage both terminological and conceptual misalignment.

3. The methodology of the Legal Taxonomy Syllabus

The main assumptions of our methodology come from studies in comparative law (Rossi and Vogel, 2004) and ontologies engineering (Klein, 2001):

- Terms –*lexical entries* for legal information–, and concepts must be distinguished; for this purpose we use lightweight ontologies (Giunchiglia and Zaihrayeu, 2007), i.e. simple taxonomic structures of primitive or composite terms together with associated definitions. They are hardly axiomatized as the intended meaning of the terms used by the community is more or less known in advance by all members, and the ontology can be limited to those structural relationships among terms that are considered as relevant (Oberle, 2005).

- We distinguish the ontology implicitly defined by EUD, the *EU level*, from the various national ontologies. Each one of these "particular" ontologies belongs to the *national level*: i.e., each national legislation refers to a distinct national legal ontology. We do not assume that the transposition of an EUD automatically introduces in a national ontology the same concepts that are present at the EU level.

- Corresponding concepts at the EU level and at the national level can be denoted by different terms in the same national language.

A standard way to properly manage large multilingual lexical databases is to do a clear distinction among terms and their interlingual acceptions (or *axies*) (Sérasset, 1994; Lyding et al., 2006).

In the LTS project to properly manage terminological and conceptual misalignment, we distinguish the notion of *legal term* from the notion of *legal concept* and we build a systematic classification based on this distinction. The basic idea in our system is that the conceptual backbone consists in a taxonomy of concepts (ontology) to which the terms can refer to express their meaning. One of the main points to keep in mind is that we do not assume the existence of a single taxonomy covering all languages. In fact, the different national systems may organize the concepts in different ways. For instance, the term *contract* corresponds to different concepts in common law and civil law, where it has the meaning of *bargain* and *agreement*, respectively (Sacco, 1999). In most complex instances, there are no homologous between terms-concepts such as *frutto civile* (legal fruit) and *income*, but respectively civil law and common law systems can achieve functionally same operational rules thanks to the functioning of the entire taxonomy of national legal concepts (Graziadei, 2004). Consequently, the LTS includes different ontologies, one for each involved national language plus one for the language of EU documents. Each language-specific ontology is related via a set of *association* links to the EU concepts, as shown in Fig. 1. Although this picture is conform to intuition, in LTS it had to be enhanced in two directions. First, it must be observed that the various national ontologies have a reference language. This is not the case for the EU ontology. For in-

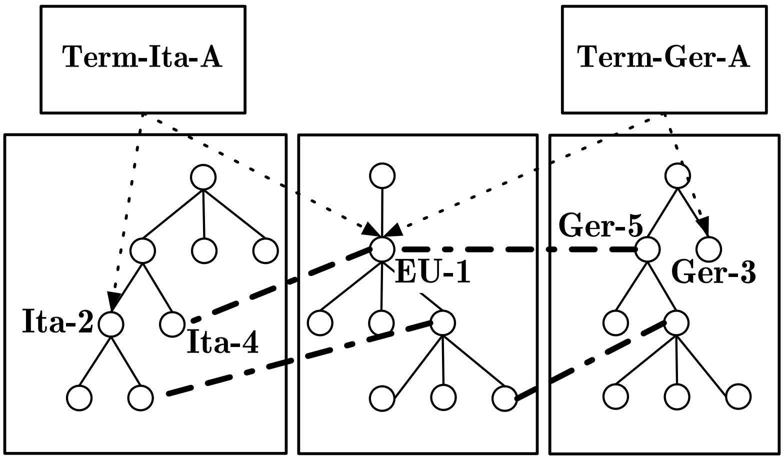


Figure 1: Relationship between ontologies and terms. The thick arcs represent the inter-ontology "association" link.

stance, a given term in English could refer either to a concept in the UK ontology or to a concept in the EU ontology. In the first case, the term is used for referring to a concept in the national UK legal system, whilst in the second one, it is used to refer to a concept used in the European directives. This is one of the main advantages of LTS. For example *klar und verständlich* could refer both to concept `Ger-379` (a concept in the German Ontology) and to concept `EU-882` (a concept in the European ontology). This is the LTS solution for facing the possibility of a correspondence only partial between the meaning of a term in the national system and the meaning of the same term in the translation of a EU directive. This feature enables the LTS to be more precise about what "translation" means. It puts at disposal a way for asserting that two terms are the translation of each other, but just in case those terms have been used in the translation of an EU directive: within LTS, we can talk about direct EU-to-national translations of terms, but only about *implicit* national-to-national translations of terms. In other words, we distinguish between *explicit* and *implicit* associations among concepts belonging to different levels. The former ones are direct links that are explicitly used by legal experts to mark a relation between concepts. The latter ones are indirect links: if we start from a concept at a given national level, by following a direct link we reach another concept at European level. Then, we will be able to see how that concept is mapped onto further concepts at the various national levels.

The situation enforced in LTS is depicted in Fig. 1, where it is represented that the Italian term *Term-Ita-A* and the German term *Term-Ger-A* have been used as corresponding terms in the translation of an EU directive, as shown by the fact that both of them refer to the same EU-concept `EU-1`. In the Italian legal system, *Term-Ita-A* has the meaning `Ita-2`. In the German legal system, *Term-Ger-A* has the meaning `Ger-3`. The EU translations of the directive is correct insofar no terms exist in Italian and German that characterize precisely the concept `EU-1` in the two languages (i.e., the "associated" concepts `Ita-4` and `Ger-5` have no corresponding legal terms). A practical example of such a situation is reported in Fig. 2, where we can see that the ontologies include different types of arcs. Beyond the usual *is-a* (linking a category to its supercategory), there are also a *purpose* arc, which relates a concept to the legal principle motivating it, and *concerns*, which refers to a general relatedness. The dotted arcs represent the reference from terms to concepts. Some terms have links both to a National ontology and to the EU Ontology (in particular, *withdrawal* vs. *recesso* and *difesa del consumatore* vs. *consumer protection*).

The last item above is especially relevant: note that this configuration of arcs specifies that: 1) *withdrawal* and *recesso* have been used as equivalent terms (concept `EU-2`) in some European Directives (e.g., Directive 90/314/EEC). 2) In that context, the term involved an act having as purpose the some kind of protection of the consumer. 3) The terms used for referring to the latter are *consumer protection* in English and *difesa del consumatore* in Italian. 4) In the British legal system, however, not all *withdrawals* have this goal, but only a subtype of them, to which the code refers to as *cancellation* (concept `Eng-3`). 5) In the Italian legal system, the term *diritto di recesso* is ambiguous, since it can be used with reference either to something concerning the *risoluzione* (concept `Ita-4`), or to something concerning the *recesso* proper (concept `Ita-3`).

The actual number of annotated terms and concepts are provided in Tables 1 and 2, respectively. Terms were initially extracted from a *corpus* of 24 EC directives, and 2 EC regulations, reported in Appendix 5.. Occurrences of such entries were detected from national transposition laws of English, French, Spanish, Italian and German jurisdictions.

Finally, it is possible to use the LTS to translate terms into different national systems via the transposed concepts at the European level, i.e. by using the implicit associations. For instance suppose that we want to translate the legal term *credito al consumo* from Italian to German. In the LTS *credito al consumo* is associated to the national umeaning `Ita-175`. We find that `Ita-175` is the transposition of the European umeaning `EU-26` (*contratto di credito*). `EU-26` is associated to the German legal term *Kreditvertrag* at European level. Again, we find that the national German transposition of `EU-26` corresponds to the

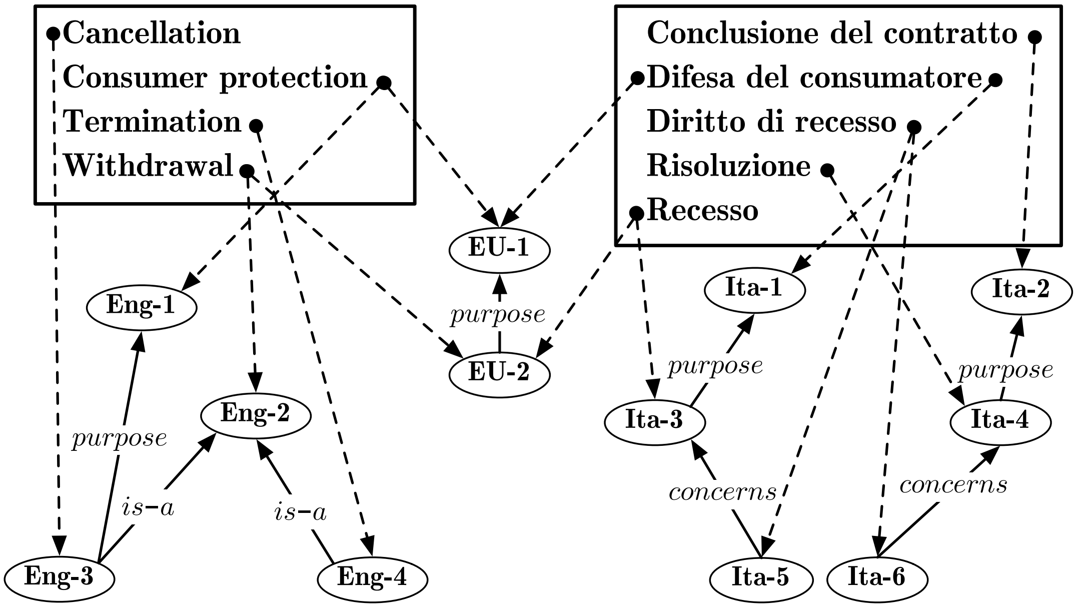


Figure 2: An example of interconnections among terms.

Table 1: Number of terms

Language	National	European
French	8	47
Italian	28	52
English	71	75
Spanish	41	60
German	66	98
total	214	332

Table 2: Number of concepts

Language	National	European
French	7	43
Italian	24	45
English	54	71
Spanish	34	56
German	52	75
total	171	290

national umeaning `Ger-32` that is associated with the national legal term *Darlehensvertrag*. Then, by using implicit links in the European ontology, we can translate the Italian legal term *credito al consumo* into the German legal term *Darlehensvertrag*.

4. Ongoing work: adding further levels

One major feature of the LTS approach relies on distinguishing legal information as belonging to different levels. At the current stage of development, the system manages terms and meanings at both EU and national levels. The former one is an ontology of legal concepts derived from the EUDs; the latter one includes national legal ontologies coming from the various national legal systems. Current approach has been devised to be general enough to account for heterogeneous legal sources (like, e.g., EUDs and "Decreti Legislativi" for European and Italian national levels respectively), and to be generalised by adding further levels. To add a new level into the system, we link a new legal ontology to that in one of the existing levels. A link is an *explicit* association connecting a concepts belonging to the new ontology and a concept belonging to the existing level ontology. We are applying this procedure in order to define an *Acquis level* to the LTS.

The EC Commission on Common Frame of Reference should provide common principles, terminology, and rules for contract law to address gaps, conflicts, and ambiguities emerging from the application of European contract law. In drafting the Action Plan the Commission emphasized that the CFR would eliminate market inefficiencies arising from the diverse implementation of European directives, providing a solution to the *non-uniform* interpretation of European contract law due to vague terms and rules, now present in the existing Acquis. In particular, two issues arise from the vague terminology of EUDs. First, directives adopt broadly defined legal concepts, therefore leaving too much discretion in their implementation to national legislators or judges. Second, directives introduce legal concepts that are different from national legal concepts. Thus, when judges face vague terms, they can either interpret them by referring to the broad principles of the acquis communautaire, or they can refer to the particular goals of the directive in question. To respond to the Action Plan, in the last few years, within the general framework of a "Network of Excellence" European Project, a research group aiming at consolidating the existing EC law is working on the "Principles of the Existing EC Private Law" or "Acquis Principles" (ACQP). These

Principles will be discussed and compared with other outcomes from different European research groups and, during a complex process of consultation with stakeholders under the direction of EC Commission, the CFR will be set up. The Acquis Principles should provide a common terminology as well as common principles to constitute a guideline for uniform implementation and interpretation of European law (Ajani and Schulte-Nölke, 2003; Schulze, 2005). One outcome of such project is the Acquis Principles glossary, i.e., a set of interconnected terms and concepts.

We introduce the *Acquis level* into the LTS by defining explicit associations between Acquis Principles concepts and EU-level concepts. For example, in Figure 3 we have that the concept `EU-25` (corresponding to the English legal term *creditor*) present in a EUD is explicitly associated to the national legal concepts `Ita-124` (*finanziatore*) and `Spa-110` (*prestamista*) for Italian and Spanish, respectively. We can add the term *creditor* from the Acquis Level by inserting an explicit association between the Acquis legal concept `AC-72`. As a consequence, the concept `AC-72` is implicitly associated to the legal concepts `Ita-124` and `Spa-110`. This fact has deep consequence on the way one can build systems for reasoning, that are allowed to make paths passing through more than two levels, thereby offering new insights (and ready-to-use associations between terms) to scholars in comparative law.

5. Conclusions

In this paper we discuss some features of the LTS, a tool for building multilingual conceptual dictionaries for the EU law. The tool is based on lightweight ontologies to allow a distinction of concepts from terms. Distinct ontologies are built at the EU level and for each national language, to deal with polysemy and terminological and conceptual misalignment.

The present work illustrates how further levels can be added to the EU and national levels. In particular, we introduced how a novel set of principles (along with a terminology) can be added to the LTS. This work has two main virtues: firstly, legal experts will be allowed to recover information from diverse kinds of data. Secondly, legal reasoning systems will benefit from a framework enriched by new explicit and implicit associations connecting Acquis, European and national levels.

Two problems arise in our approach: the first one is theoretical, and it concerns the issue of evaluating the system. Secondly, the amount of work needed to annotate the EUDs with concepts, terms and their transpositions, is huge. Future work will involve exploring ways to extend the LTS ontology and populating it at the various levels by semi-automatic approaches.

6. References

G. Ajani and H. Schulte-Nölke. 2003. The Action Plan on a More Coherent European Contract Law: Response on Behalf of the Acquis Group.

Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Alessandro Mazzei, and Piercarlo Rossi. 2007a. Multilingual Ontological Analysis of European Directives. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 21–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Gianmaria Ajani, Leonardo Lesmo, Guido Boella, Alessandro Mazzei, and Piercarlo Rossi. 2007b. Terminological and ontological analysis of european directives: multilinguism in law. In *ICAIL*, pages 43–48.

A. Boer, T.M. van Engers, and R. Winkels. 2003. Using ontologies for comparing and harmonizing legislation. In *ICAIL*, pages 60–69.

P. Casanovas, N. Casellas, C. Tempich, D. Vrandecic, and R. Benjamins. 2005. OPJK modeling methodology. In *Proceedings of the ICAIL Workshop: LOAIT 2005*.

S. Després and S. Szulman. February 2007. Merging of legal micro-ontologies from european directives. *Journal of Artificial Intelligence and Law*.

European Commission. 2003. Communication from the Commission to the European Parliament and the Council - A More Coherent European Contract Law - An Action Plan. COM.

Emmanuel Giguet and Pierre-Sylvain Luquet. 2006. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 271–278, Sydney, Australia, July. Association for Computational Linguistics.

Fausto Giunchiglia and Ilya Zaihrayeu. 2007. Lightweight Ontologies. Technical Report DIT-07-071, University of Trento, Department of Information and Communication Technology, 38050 Povo – Trento (Italy), Via Sommarive 14, October.

M. Graziadei. 2004. Tuttifrutti. In P. Birks and A. Pretto, editors, *Themes in Comparative Law*, pages –. Oxford University Press.

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

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 Language Resources and Interoperability, ACL06*, pages 25–31.

D. Oberle, editor. 2005. *Semantic Management of Middleware*. Springer Science+Business and Media.

Research Group on the Existing EC Private Law. 2007. *Principles of the Existing EC Contract Law (Acquis Principles) Contract I*. Sellier, European Law Publishers.

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.

R. Sacco. 1999. Contract. *European Review of Private Law*, 2:237–240.

R. Schulze. 2005. European Private Law and Existing EC Law. *European Review of Private Law (ERPL)*.

G. Sérasset. 1994. Interlingual lexical organization for

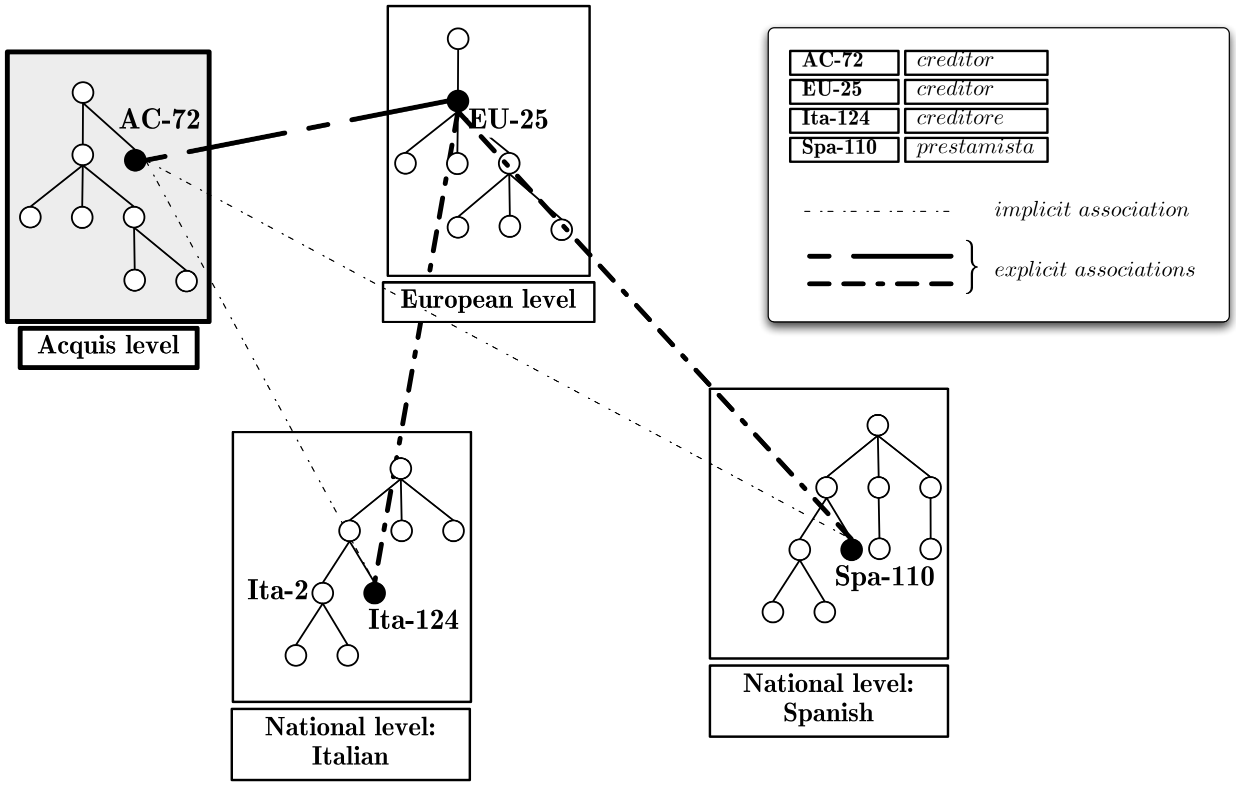


Figure 3: LTS augmented with the Acquis level. Thick lines indicate *explicit* associations; thin lines indicate *implicit* associations.

multilingual lexical databases in NADIA. In *Proc. COLING94*, pages 278–282.

Daniela Tiscornia. 2007. The Lois Project: Lexical Ontologies for Legal Information . In *Proceedings of the V Legislative XML Workshop*. Association for Computational Linguistics.

P. Vossen, W. Peters, and J. Gonzalo. 1999. Towards a universal index of meaning. In *Proc. ACL-99 Siglex Workshop*.

APPENDIX A: List of EC directives

Core directives

- 84/450/EEC concerning misleading advertising
- 85/374/EEC concerning liability for defective products
- 85/577/EEC to protect the consumer in respect of contracts negotiated away from business premises
- 87/102/EEC concerning consumer credit
- 90/88 concerning consumer credit
- 90/314/EEC on package travel, package holidays and package tours
- 93/13/EEC on unfair terms in consumer contracts
- 94/47/EC on the protection of purchasers in respect of certain aspects of contracts relating to the purchase of the right to use immovable properties on a timeshare basis
- 97/7/EC on the protection of consumers in respect of distance contracts
- 97/55/EC concerning misleading advertising so as to include comparative advertising
- 98/6 on consumer protection in the indication of the prices of products offered to consumers
- 98/7 concerning consumer credit
- 98/27/EC on injunctions for the protection of consumers' interests
- 99/44/EC on certain aspects of the sale of consumer goods and associated guarantees
- 2000/13/EC relating to labelling, presentation and advertising of foodstuff
- 2001/95 on general product safety
- 2002/65/EC concerning the distance marketing of consumer financial services
- Regulation 2006/2004/EC on co-operation between national authorities responsible for the enforcement of consumer protection laws
- Directive 2005/29/EC on Unfair Commercial Practices

Ancillary directives

- 76/768/EEC relating to cosmetic products
- 88/378/EEC on toy safety
- 89/552/EEC on TV broadcasting activities
- 96/74/EC on textile names
- 97/5/EC on cross border credit transfers
- Recommendation 98/257 on the principles applicable to bodies responsible for the out-of-court settlement of consumer disputes
- 2000/31/EC on electronic commerce
- Regulation 2560/2001/EC on cross-border payments in Euro

Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain

Eneldo Loza Mencía and Johannes Fürnkranz

Knowledge Engineering Group
Technische Universität Darmstadt
[eneldo,juffi]@ke.informatik.tu-darmstadt.de

Abstract

In this paper we evaluate the performance of multilabel classification algorithms on the EUR-Lex database of legal documents of the European Union. On the same set of underlying documents, we defined three different large-scale multilabel problems with up to 4000 classes. On these datasets, we compared three algorithms: (i) the well-known one-against-all approach (OAA); (ii) the multiclass multilabel perceptron algorithm (MMP), which modifies the OAA ensemble by respecting dependencies between the base classifiers in the training protocol of the classifier ensemble; and (iii) the multilabel pairwise perceptron algorithm (MLPP), which unlike the previous algorithms trains one base classifier for each pair of classes. All algorithms use the simple but very efficient perceptron algorithm as the underlying classifier. This makes them very suitable for large-scale multilabel classification problems. While previous work has already shown that the latter approach outperforms the other two approaches in terms of predictive accuracy, its key problem is that it has to store one classifier for each pair of classes. The key contribution of this work is to demonstrate a novel technique that makes the pairwise approach feasible for problems with large number of classes, such as those studied in this work. Our results on the EUR-Lex database illustrate the effectiveness of the pairwise approach and the efficiency of the MMP algorithm. We also show that it is feasible to efficiently and effectively handle very large multilabel problems.

1. Introduction

Recently, *multilabel classification* problems, where the task is to associate an object with an unbounded set of classes instead of exactly one, have received increased attention in the literature. As a consequence, new algorithms have been developed or adapted to automatically solve the task of multilabel classification. But simultaneously an increased number of new scenarios have been identified and higher demands are continuously made to the existing algorithms. This concerns not only challenges due to large scale instance spaces, large numbers of instances and numbers of features, but particularly due to the number of possible classes. In particular in text classification, these type of problems are very common. The number of possible categories that can typically be assigned to each document varies from a few dozen to several hundred. In this paper, we study a challenging new domain, namely assigning documents of the EUR-Lex database to a few of $\approx 4,000$ possible labels. The EUR-Lex database is a freely accessible document management system for legal documents of the European Union. We chose this database for several reasons:

- it contains multiple classifications of the same documents, making it possible to analyze the effects of different classification properties using the same underlying reference data without resorting to artificial or manipulated classifications,
- the overwhelming number of produced documents make the legal domain a very attractive field for employing supportive automated solutions and therefore a machine learning scenario in step with actual practice,
- the documents are available in several European languages and are hence very interesting e.g. for the wide field of multi- and cross-lingual text classification,
- and, finally, the data is freely accessible.

The simplest strategy to tackle the multilabel problem with existing techniques is to use *one-against-all* binarization, in the multilabel setting also referred to as the *binary relevance* method. It decomposes the original problem into less complex, binary problems, by learning one classifier for each class, using all objects of this class as positive examples and all other objects as negative examples. At query time, each binary classifier predicts whether its class is relevant for the query example or not, resulting in a set of relevant labels. While this technique can potentially be used to transform any binary classifier into a multilabel classifier and it is often used in practical applications, the question remains whether this general approach can fully adapt to the particular needs of multilabel classification, because it trains each class independently of all other classes.

A recently proposed alternative that tries to tackle this problem is the *multilabel multiclass perceptron* (MMP) algorithm developed by Crammer and Singer (2003), which adapts the one-against-all approach to the multilabel case. Instead of learning the relevance of each class individually and independently, MMP incrementally trains the entire classifier ensemble as a whole so that it predicts a real-valued relevance value for each class. This is done by always evaluating the performance of the entire ensemble, and only producing training examples for the individual classifiers when their corresponding classes are misplaced in the ranking. It uses perceptrons as base classifiers, because they are simple and efficient, and because for high-dimensional problems such as text classification, linear discriminants are sufficiently expressive.

An alternative training method for an effective ensemble of perceptrons is the pairwise decomposition of the initial problem proposed by the *multilabel pairwise perceptron* (MLPP) algorithm (Loza Mencía and Fürnkranz, 2008). In this method, one perceptron is trained for each possible class pair, using the examples belonging to the two classes as positive or negative examples respectively. During prediction, an overall ranking of the classes is determined by combining the predictions of the individual base perceptrons, e.g. by voting. The first main advantage of the pairwise approach is its effectiveness: decomposing the problem into smaller subproblem will yield simpler, often linear decision boundaries, and this usually leads to an increased prediction performance (Knerr et al., 1992; Hsu and Lin, 2002; Fürnkranz, 2002). However, the second advantage of increased efficiency in training compared to one-against-all ensembles (Fürnkranz, 2002) is counteracted by the quadratic number of base classifiers required in proportion to the number of classes. While the pairwise decomposition in combination with perceptrons as base classifiers is well applicable for a large feature space and a large amount of training instances (Loza Mencía and Fürnkranz, 2008), the large number of classes in the EUR-Lex database constitutes an insurmountable obstacle: in the most complex setting approximately 8,000,000 perceptrons would be needed to represent a pairwise ensemble.

We will therefore introduce and analyze a novel variant that addresses this problem by representing the perceptron in its dual form, i.e. the perceptrons are formulated as a combination of the documents that were used during training instead of explicitly as a linear hyperplane. This reduces the dependence on the number of classes and therefore allows the *Dual MLPP* algorithm to handle the tasks in the EUR-Lex database.

The MMP algorithm has already been used on large scale data sets such as the Reuters Corpus Volume 1 (Lewis et al., 2004) with its over 800,000 examples and approx. 100 classes (Crammer and Singer, 2003). In the experiments the MMP algorithm was able to substantially improve the performance of the one-against-all method with perceptrons as base classifiers. In a recent evaluation on the same data the pairwise approach showed an even higher performance than the MMP algorithm (Loza Mencía and Fürnkranz, 2008), demonstrating the applicability of MLPP to large-scale problems. In this paper we will analyze if these results can be repeated with the EUR-Lex database in three different settings, one with approx. 200, another with 400 and finally one with 4000 possible classes. Note that the latter problem has more classes by an order of magnitude than other known applications of these algorithms.

A shortcoming of MMP and the pairwise method is that the resulting prediction is not any more a set of classes as expected for a multilabel task, but a ranking of class relevance scores. However, it is possible to obtain the desired output in an additional step that selects classes which exceed a determined relevance value. Different methods exist for determining the threshold, a good overview can be found in Sebastiani (2002). Recently, Brinker et al. (2006) introduced the idea of using an artificial label that encodes the boundary between relevant and irrelevant labels for each example. In this paper, we will concentrate on the label ranking task, which also enables a more detailed evaluation of the classification performance.

2. Preliminaries

We represent an instance or object as a vector $\bar{\mathrm{x}} = (x_1, \ldots, x_M)$ in a feature space $\mathcal{X} \subseteq \mathbb{R}^N$. Each instance $\bar{\mathrm{x}}_i$ is assigned to a set of relevant labels $\mathcal{y}_i$, a subset of the K possible classes $\mathcal{Y} = \{c_1, \ldots, c_K\}$. For multilabel problems, the cardinality $|\mathcal{y}_i|$ of the label sets is not restricted, whereas for binary problems $|\mathcal{y}_i| = 1$. For the sake of simplicity we use the following notation for the binary case: we define $\mathcal{Y} = \{1, -1\}$ as the set of classes so that each object $\bar{\mathrm{x}}_i$ is assigned to a $y_i \in \{1, -1\}$, $\mathcal{y}_i = \{y_i\}$.

2.1. Ranking Loss Functions

In order to evaluate the predicted ranking we use different *ranking losses*. The losses are computed comparing the ranking with the true set of relevant classes, each of them focusing on different aspects. For a given instance $\bar{\mathrm{x}}$, a relevant label set $\mathcal{y}$, a negative label set $\overline{\mathcal{y}} = \mathcal{Y}\backslash\mathcal{y}$ and a given predicted ranking $r : \mathcal{Y} \rightarrow \{1 \ldots K\}$, with $r(c)$ returning the position of class c in the ranking, the different loss functions are computed as follows:

IsErr The is-error loss determines whether $r(c) < r(c')$ for all relevant classes $c \in \mathcal{y}$ and all irrelevant classes $c' \in \overline{\mathcal{y}}$. It returns 0 for a completely correct, *perfect ranking*, and 1 for an incorrect ranking, irrespective of 'how wrong' the ranking is.

OneErr The one error loss is 1 if the top class in the ranking is not a relevant class, otherwise 0 if the top class is relevant, independently of the positions of the remaining relevant classes.

RankLoss The ranking loss returns the number of pairs of labels which are not correctly ordered normalized by the total number of possible pairs. As IsErr, it is 0 for a perfect ranking, but it additionally differentiates between different degrees of errors.

$$E \overset{\text{def}}{=} \{(c, c') \mid r(c) > r(c')\} \subseteq \mathcal{y} \times \overline{\mathcal{y}} \qquad (1)$$

$$\delta_{\text{RankLoss}} \overset{\text{def}}{=} \frac{|E|}{|\mathcal{y}||\overline{\mathcal{y}}|} \qquad (2)$$

Margin The margin loss returns the number of positions between the worst ranked positive and the best ranked negative classes. This is directly related to the number of wrongly ranked classes, i.e. the positive classes that are ordered below a negative class, or vice versa. We denote this set by F.

$$\begin{aligned} F \overset{\text{def}}{=} &\{c \in \mathcal{y} \mid r(c) > r(c'), c' \in \overline{\mathcal{y}}\} \\ &\cup \{c' \in \overline{\mathcal{y}} \mid r(c) > r(c'), c \in \mathcal{y}\} \end{aligned} \qquad (3)$$

$$\begin{aligned} \delta_{\text{Margin}} \overset{\text{def}}{=} &\max(0, \max\{r(c) \mid c \in \mathcal{y}\} \\ &- \min\{r(c') \mid c' \notin \mathcal{y}\}) \end{aligned} \qquad (4)$$

AvgP Average Precision is commonly used in Information Retrieval and computes for each relevant label the percentage of relevant labels among all labels that are ranked before it, and averages these percentages over all relevant labels. In order to bring this loss in line with the others so that an optimal ranking is 0, we revert the measure.

$$\delta_{\text{AVGP}} \overset{\text{def}}{=} 1 - \frac{1}{\mathcal{y}} \sum_{c \in \mathcal{y}} \frac{|\{c^* \in \mathcal{y} \mid r(c^*) \leq r(c)\}|}{r(c)} \tag{5}$$

2.2. Perceptrons

We use the simple but fast perceptrons as base classifiers (Rosenblatt, 1958). As Support Vector Machines (SVM), their decision function describes a hyperplane that divides the N-dimensional space into two halves corresponding to positive and negative examples. We use a version that works without learning rate and threshold:

$$o'(\bar{x}) = sgn(\bar{x} \cdot \bar{w}) \tag{6}$$

with the internal weight vector $\bar{w}$ and $sgn(t) = 1$ for $t \geq 0$ and -1 otherwise. Two sets of points are called *linearly separable* if there exists a *separating hyperplane* between them. If this is the case and the examples are seen iteratively, the following update rule provably finds a separating hyperplane (cf., e.g., (Bishop, 1995)).

$$\alpha_i = (y_i - o'(\bar{x}_i)) \qquad \bar{w}_{i+1} = \bar{w}_i + \alpha_i \bar{x}_i \tag{7}$$

It is important to see that the final weight vector can also be represented as linear combination of the training examples:

$$\bar{w} = \sum_{i=1}^{M} \alpha_i \bar{x}_i \qquad o'(\bar{x}) = sgn(\sum_{i=1}^{M} \alpha_i \cdot \bar{x}_i \bar{x}) \tag{8}$$

assuming M as number of seen training examples and $\alpha_i \in \{-1, 0, 1\}$. The perceptron can hence be coded implicitly as a vector of instance weights $\alpha = (\alpha_1, \dots, \alpha_M)$ instead of explicitly as a vector of feature weights. This representation is denominated the dual form and is crucial for developing the memory efficient variant in section . The main reason for choosing Perceptron as our base classifier is because, contrary to SVMs, they can be trained efficiently in an incremental setting, which makes them particularly well-suited for large-scale classification problems such as the RCV1 benchmark (Lewis et al., 2004), without forfeiting too much accuracy though SVMs find the *maximum-margin hyperplane* (Freund and Schapire, 1999; Crammer and Singer, 2003; Shalev-Shwartz and Singer, 2005).

2.3. Binary Relevance Ranking

In the binary relevance (BR) or one-against-all (OAA) method, a multilabel training set with K possible classes is decomposed into K binary training sets of the same size that are then used to train K binary classifiers. So for each pair $(\bar{x}_i, \mathcal{y}_i)$ in the original training set K different pairs of instances and binary class assignments $(\bar{x}_i, y_{i_j})$ with $j = 1 \dots K$ are generated as follows:

$$y_{i_j} = \begin{cases} 1 & c_j \in \mathcal{y}_i \\ -1 & otherwise \end{cases} \tag{9}$$

Supposing we use perceptrons as base learners, K different o'_j classifier are trained in order to determine the relevance of c_j. In consequence, the combined prediction of the binary relevance classifier for an instance $\bar{x}$ would be the set $\{c_j \mid o'_j(\bar{x}) = 1\}$. If, in contrast, we want to obtain a ranking of classes according to their relevance, we can simply use the result of the internal computation of the perceptrons as a measure of relevance. According to Equation 6 the desired linear combination is the inner product $o_j(\bar{x}) = \bar{x} \cdot \bar{w}_j$ (ignoring ω as mentioned above). So the result of the prediction is a vector $\bar{o}(\bar{x}) = (\bar{x}\bar{w}_1, \dots, \bar{x}\bar{w}_K)$ where component j corresponds to the relevance of class c_j. Ties are broken randomly to not favor any particular class.

```
Require: Training example pair $(\bar{x}, \mathcal{y})$, perceptrons $\bar{w}_1, \dots, \bar{w}_K$
calculate $\bar{x}\bar{w}_1, \dots, \bar{x}\bar{w}_K$, loss $\delta$
if $\delta > 0$ then                                  ▷ only if ranking is not perfect
    calculate error sets $E$, $F$
    for each $c \in F$ do $\tau_c \leftarrow 0$                    ▷ initialize $\tau$'s
    for each $(c, c') \in E$ do
        $p \leftarrow$ PENALTY($\bar{x}\bar{w}_1, \dots, \bar{x}\bar{w}_K$)
        $\tau_c \leftarrow \tau_c + p$                          ▷ push up pos. classes
        $\tau_{c'} \leftarrow \tau_{c'} - p$                        ▷ push down neg. classes
        $\sigma \leftarrow \sigma + p$                            ▷ for normalization
    normalize $\tau$'s
    for each $c \in F$ do
        $\bar{w}_c \leftarrow \bar{w}_c + \delta \frac{\tau_c}{\sigma} \cdot \bar{x}$                ▷ update perceptrons
return $\bar{w}_1 \dots \bar{w}_K$                      ▷ return updated perceptrons
```

Figure 1: Pseudocode of the training method of the MMP algorithm

2.4. Multiclass Multilabel Perceptrons

MMPs were proposed as an extension of the one-against-all algorithm with perceptrons as base learners (Crammer and Singer, 2003). Just as in binary relevance, one perceptron is trained for each class, and the prediction is calculated via the inner products. The difference lies in the update method: while in the binary relevance method all perceptrons are trained independently to return a value greater or smaller than zero, depending on the relevance of the classes for a certain instance, MMPs are trained to produce a good ranking so that the relevant classes are all ranked above the irrelevant classes. The perceptrons therefore cannot be trained independently, considering that the target value for each perceptron depends strongly on the values returned by the other perceptrons.

The pseudocode in Fig. 1 describes the MMP training algorithm. When the MMP algorithm receives a training instance $\bar{x}$, it calculates the inner products, the ranking and the loss on this ranking in order to determine whether the current model needs an update. For determining the ranking loss, any of the methods of Sec. 2.1. is appropriate, since they all return a low value on good rankings. If the ranking is perfect, the algorithm is done, otherwise it calculates the error set of wrongly ordered class pairs E. The wrongly ranked classes are also represented in F. In the next step, each class that is present in a pair of E receives a penalty score. This is done according to a selectable penalty function, being the uniform update method, where each pair in E receives the same score, the most successful one (Crammer and Singer, 2003). In the next step, the update weights τ are

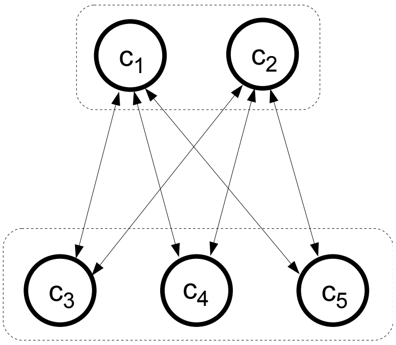


Figure 2: MLPP training: training example $\bar{x}$ belongs to $y = \{c_1, c_2\}$, $\bar{y} = \{c_3, c_4, c_5\}$ are the irrelevant classes, the arrows represent the trained perceptrons.

```
Require: Training example pair $(\bar{x}, y)$,
    perceptrons $\{\bar{w}_{u,v} \mid u < v, c_u, c_v \in \mathcal{Y}\}$
for each $(c_u, c_v) \in y \times \bar{y}$ do
    if $u < v$ then
        $\bar{w}_{u,v} \leftarrow$ TRAINPERCEPTRON($\bar{w}_{u,v}, (\bar{x}, 1)$)
                                        ▷ train as positive example
    else
        $\bar{w}_{v,u} \leftarrow$ TRAINPERCEPTRON($\bar{w}_{v,u}, (\bar{x}, -1)$)
                                        ▷ train as negative example
return $\{\bar{w}_{u,v} \mid u < v, c_u, c_v \in \mathcal{Y}\}$
                                        ▷ updated perceptrons
```

Figure 3: Pseudocode of the training method of the MLPP algorithm.

normalized and each perceptron whose class was wrongly ordered is updated.

2.5. Multilabel Pairwise Perceptrons

In the pairwise binarization method, one classifier is trained for each pair of classes, i.e., a problem with K different classes is decomposed into $\frac{K(K-1)}{2}$ smaller subproblems. For each pair of classes (c_u, c_v), only examples belonging to either c_u or c_v are used to train the corresponding classifier $o'_{u,v}$. All other examples are ignored. In the multilabel case, an example is added to the training set for classifier $o'_{u,v}$ if u is a relevant class and v is an irrelevant class, i.e., $(u, v) \in y \times \bar{y}$ (cf. Figure 2). We will typically assume $u < v$, and training examples of class u will receive a training signal of $+1$, whereas training examples of class v will be classified with -1. Figure 3 shows the training algorithm in pseudocode. Of course MLPPs can also be trained incrementally.

In order to return a class ranking we use a simple voting strategy, known as *max-wins*. Given a test instance, each perceptron delivers a prediction for one of its two classes. This prediction is decoded into a vote for this particular class. After the evaluation of all $\frac{K(K-1)}{2}$ perceptrons the classes are ordered according to their sum of votes.[1]

Figure 4 shows a possible result of classifying the sample instance of Figure 2. Perceptron $o'_{1,5}$ predicts (correctly) the first class, consequently c_1 receives one vote and class c_5 zero (denoted by $o'_{1,5} = 1$ in the first and $o'_{5,1} = -1$ in the last row). All 10 perceptrons (the values in the upper right corner can be deduced due to the symmetry property of the perceptrons) are evaluated though only six are 'qualified' since they were trained with the original example.

$o'_{1,2} = 1$	$o'_{2,1} =$ -1	$o'_{3,1} =$ -1	$o'_{4,1} =$ -1	$o'_{5,1} =$ -1
$o'_{1,3} = 1$	$o'_{2,3} = 1$	$o'_{3,2} =$ -1	$o'_{4,2} =$ -1	$o'_{5,2} =$ -1
$o'_{1,4} = 1$	$o'_{2,4} = 1$	$o'_{3,4} = 1$	$o'_{4,3} =$ -1	$o'_{5,3} =$ -1
$o'_{1,5} = 1$	$o'_{2,5} = 1$	$o'_{3,5} = 1$	$o'_{4,5} = 1$	$o'_{5,4} =$ -1
$v_1 = 4$	$v_2 = 3$	$v_3 = 2$	$v_4 = 1$	$v_5 = 0$

Figure 4: MLPP voting: an example $\bar{x}$ is classified by all 10 base perceptrons $o'_{u,v}, u \neq v,\ c_u, c_v \in \mathcal{Y}$. Note the redundancy given by $o'_{u,v} = -o'_{v,u}$. The last line counts the positive outcomes for each class.

This may be disturbing at first sight since many 'unqualified' perceptrons are involved in the voting process: $o'_{1,2}$ is asked for instance though it cannot know anything relevant in order to determine if $\bar{x}$ belongs to c_1 or c_2 since it was neither trained on this example nor on other examples belonging simultaneously to both classes (or to none of both). In the worst case the noisy votes concentrate on a single negative classes, which would lead to misclassifications. But note that any class can at most receive $K-1$ votes, so that in the extreme case when the qualified perceptrons all classify correctly and the unqualified ones concentrate on a single class, a positive class would still receive at least $K - |y|$ and a negative at most $K - |y| - 1$ votes. Class c_3 in Figure 4 is an example for this: It receives all possible noisy votes but still loses against the positive classes c_1 and c_2.

The pairwise binarization method is often regarded as superior to binary relevance because it profits from simpler decision boundaries in the subproblems (Fürnkranz, 2002; Hsu and Lin, 2002). In the case of an equal class distribution, the subproblems have $\frac{2}{K}$ times the original size whereas binary relevance maintains the size. Typically, this goes hand in hand with an increase of the space where a separating hyperplane can be found. Particularly in the case of text classification the obtained benefit clearly exists. An evaluation of the pairwise approach on the Reuters Corpus Volume 1 (Lewis et al., 2004), which contains over 100 classes and 800,000 documents, showed a significant and substantial improvement over the MMP method (Loza Mencía and Fürnkranz, 2008). This encourages us to apply the pairwise decomposition to the EUR-Lex database, with the main obstacle of the quadratic number of base classifier in relationship to the number of classes. Since this problem can not be coped for the present classifications in EUR-Lex, we propose to reformulate the MLPP algorithm in the way described in the next section.

3. Dual Multilabel Pairwise Perceptrons

With an increasing number of classes the required memory by the MLPP algorithm grows quadratically and even on modern computers with a huge amount of memory this problem becomes unsolvable for a high number of classes. For the EUROVOC classification, the use of MLPP would mean maintaining approximately 8,000,000 perceptrons in memory. In order to circumvent this obstacle we

[1] Ties are broken randomly in our case.

reformulate the MLPP ensemble of perceptrons in dual form as we did with one single perceptron in Equation 8. In contrast to MLPP, the training examples are thus required and have to be kept in memory in addition to the associated weights, as a base perceptron is now represented as $\bar{w}_{u,v} = \sum_{i=1}^{M} \alpha^t_{u,v}\bar{x}_i$. This makes an additional loop over the training examples inevitable every time a prediction is demanded. But fortunately it is not necessary to recompute all $\bar{x}_i\bar{x}$ for each base perceptron since we can reuse them by iterating over the training examples in the outer loop, as can be seen in the following equations:

$$\begin{aligned}
\bar{w}_{1,2}\bar{x} &= \alpha^1_{1,2}\bar{x}_1\bar{x} + \alpha^2_{1,2}\bar{x}_2\bar{x} + \ldots + \alpha^M_{1,2}\bar{x}_M\bar{x} \\
\bar{w}_{1,3}\bar{x} &= \alpha^1_{1,3}\bar{x}_1\bar{x} + \alpha^2_{1,3}\bar{x}_2\bar{x} + \ldots + \alpha^M_{1,3}\bar{x}_M\bar{x} \\
&\vdots \\
\bar{w}_{1,K}\bar{x} &= \alpha^1_{1,K}\bar{x}_1\bar{x} + \alpha^2_{1,K}\bar{x}_2\bar{x} + \ldots + \alpha^M_{1,K}\bar{x}_M\bar{x} \\
\bar{w}_{2,3}\bar{x} &= \alpha^1_{2,3}\bar{x}_1\bar{x} + \alpha^2_{2,3}\bar{x}_2\bar{x} + \ldots + \alpha^M_{2,3}\bar{x}_M\bar{x} \\
&\vdots
\end{aligned} \tag{10}$$

By advancing column by column it is not necessary to repeat the dot products computations, however it is necessary to store the intermediate values, as can also be seen in the pseudocode of the training and prediction phases in Figures 5 and 6. We denote this variant of training the pairwise perceptrons the dual multilabel pairwise perceptrons algorithm (DMLPP).

Despite the consequences for the memory requirements and the run-time analyzed in detail in Section 4., the dual representation allows for using the kernel trick, i.e. to replace the dot product by a kernel function, in order to be able to solve originally not linearly separable problems. However, this is not necessary in our case since text problems are in general linearly separable.

Note also that the pseudocode needs to be slightly adapted when the DMLPP algorithm is trained in more than one epoch, i.e. the training set is presented to the learning algorithm more than once. It is sufficient to modify the assignment in line 8 in Figure 5 to an additive update $\alpha^t_{u,v} = \alpha^t_{u,v} + 1$ for a revisited example $\bar{x}_t$. This setting is particularly interesting for the dual variant since, when the training set is not too big, memorizing the inner products can boost the subsequent epochs in a substantial way, making the algorithm interesting even if the number of classes is small.

4. Computational Complexity

The notation used in this section is the following: K denotes the number of possible classes, L the average number of relevant classes per instance in the training set, N the number of attributes and N' the average number of attributes not zero (size of the sparse representation of an instance), and M denotes the size of the training set. For each complexity we will give an upper bound O in Landau notation. We will indicate the runtime complexity in terms of real value additions and multiplications ignoring operations that have to be performed by all algorithms such as sorting or internal real value operations. Additionally, we will present the complexities per instance as all algorithms are incrementally trainable. We will also concentrate on the comparison between MLPP and the implicit representation DMLPP.

```
Require: New training example pair $(\bar{x}_M, y)$,
    training examples $\bar{x}_1 \ldots \bar{x}_{M-1}$,
    weights $\{\alpha^i_{u,v} \mid c_u, c_v \in \mathcal{Y}, 0 < i < M\}$
1: for each $\bar{x}_i \in \bar{x}_1 \ldots \bar{x}_{M-1}$ do
2:     $p_i \leftarrow \bar{x}_i \cdot \bar{x}_M$
3:     for each $(c_u, c_v) \in y \times \bar{y}$ do
4:         if $\alpha^i_{u,v} \neq 0$ then
5:             $s_{u,v} \leftarrow s_{u,v} + \alpha^i_{u,v} \cdot p_t$
                                          ▷ note that $s_{u,v} = -s_{v,u}$
6: for each $(c_u, c_v) \in y \times \bar{y}$ do
7:     if $s_{u,v} < 0$ then
8:         $\alpha^M_{u,v} \leftarrow 1$              ▷ note that $\alpha_{u,v} = -\alpha_{v,u}$
9: return $\{\alpha^M_{u,v} \mid (c_u, c_v) \in y \times \bar{y}\}$      ▷ return new weights
```

Figure 5: Pseudocode of the training method of the DMLPP algorithm.

```
Require: example $\bar{x}$ for classification,
    training examples $\bar{x}_1 \ldots \bar{x}_{M-1}$,
    weights $\{\alpha^i_{u,v} \mid c_u, c_v \in \mathcal{Y}, 0 < i < M\}$
for each $\bar{x}_i \in \bar{x}_1 \ldots \bar{x}_M$ do
    $p \leftarrow \bar{x}_i \cdot \bar{x}$
    for each $(c_u, c_v) \in y_i \times \bar{y}_t$ do
        if $\alpha^i_{u,v} \neq 0$ then
            $s_{u,v} \leftarrow s_{u,v} + \alpha^i_{u,v} \cdot p$
for each $(c_u, c_v) \in \mathcal{Y} \times \mathcal{Y}$ do
    if $u \neq v \vee s_{u,v} > 0$ then
        $v_u \leftarrow v_u + 1$
return voting $\bar{v} = (v_1, \ldots, v_{|y|})$          ▷ return voting
```

Figure 6: Pseudocode of the prediction phase of the DMLPP algorithm.

The MLPP algorithm has to keep $\frac{K(K-1)}{2}$ perceptrons, each with N weights in memory, hence we need $O(K^2N)$ memory. The DMLPP algorithms keeps the whole training set in memory, and additionally requires for each training example $\bar{x}$ access to the weights of all class pairs $y \times \bar{y}$. Furthermore, it has to intermediately store the resulting scores for each base perceptron during prediction, hence the complexity is $O(MLK + MN' + K^2) = O(M(LK + N') + K^2)$.[2] We can see that MLPP is applicable especially if the number of classes is low and the number of examples high, whereas DMLPP is suitable when the number of classes is high, however it does not handle huge training sets very well.

For processing one training example, $O(LK)$ dot products have to be computed by MLPP, one for each associated perceptron. Assuming that a dot product computation costs $O(N')$, we obtain a complexity of $O(LKN')$ per trai-

[2] Note that we do not estimate L as $O(K)$ since both values are not of the same order of magnitude in practice. For the same reason we distinguish between N and N' since particularly in text classification both values are not linked: a text document often turns out to employ around 100 different words whereas the size of the vocabulary of a the whole corpus can easily reach 100,000 words (although this number is normally reduced by feature selection).

	training time	testing time	memory requirement
MMP, BR	$O(KN')$	$O(KN')$	$O(KN)$
MLPP	$O(LKN')$	$O(K^2N')$	$O(K^2N)$
DMLPP	$O(M(LK + N'))$	$O(M(LK + N'))$	$O(M(LK + N') + K^2)$

Table 1: Computational complexity given in expected number of addition and multiplication operations. K: *#classes*, L: *avg. #labels per instance*, M: *#training examples*, N: *#attributes*, N': *#attributes*$\neq 0$, $\hat{\delta}$: *avg. Loss*, $\hat{\delta}_{per}, \hat{\delta}_{\textsc{IsErr}} \leq 1, \hat{\delta}_{\textsc{Margin}} < K$.

ning example. Similarly, the DMLPP spends M dot product computations. In addition the summation of the scores costs $O(LK)$ per training instance, leading to $O(M(LK + N'))$ operations. It is obvious that MLPP has a clear advantage over DMLPP in terms of training time, unless K is of the order of magnitude of M or the model is trained over several epochs, as already outlined in the previous Section 3.

During prediction the MLPP evaluates all perceptrons, leading to $O(K^2N')$ computations. The dual variant again iterates over all training examples and associated weights, hence the complexity is $O(M(LK + N'))$. At this phase DMLPP benefits from the linear dependence of the number of classes in contrast to the quadratic relationship of the MLPP. Roughly speaking the breaking point when DMLPP is faster in prediction is approximately when the square of the number of classes is clearly greater than the number of training documents. We can find a similar trade-off for the memory requirements with the difference that the factor between sparse and total number of attributes becomes more important, leading earlier to the breaking point when the sparseness is high.

A compilation of the analysis can be found in Table 1, together with the complexities of MMP and BR. A more detailed comparison between MMP and MLPP is available from Loza Mencía and Fürnkranz (2008).

In summary, it can be stated that the dual form of the MLPP balances the relationship between training and prediction time by increasing training and decreasing prediction costs, and especially benefits from a decreased prediction time and memory savings when the number of classes is large, which was the main obstacle to applying the pairwise approach to large scale problems in terms of classes.

5. EUR-Lex Repository

The EUR-Lex/CELEX (Communitatis Europeae LEX) Site[3] provides a freely accessible repository for European Union law texts. The documents include the official Journal of the European Union, treaties, international agreements, legislation in force, legislation in preparation, case-law and parliamentary questions. The documents are available in most of the languages of the EU, and in the HTML and PDF formats. We retrieved the HTML versions with bibliographic notes recursively from all (non empty) documents in the English version of the *Directory of Community legislation in force*[4], in total 19,596 documents. Only documents related to secondary law (in contrast to primary law, the constitutional treaties of the European Union) and international agreements are included in this repository. The legal form of the included acts are mostly *decisions* (8,917 documents), *regulations* (5,706), *directives* (1,898) and *agreements* (1,597).

The bibliographic notes of the documents contain information such as dates of effect and validity, authors, relationships to other documents and classifications. The classifications include the assignment to several EUROVOC descriptors, directory codes and subject matters, hence all classifications are multilabel ones. EUROVOC is a multilingual thesaurus providing a controlled vocabulary for European Institutions[5]. Documents in the documentation systems of the EU are indexed using this thesaurus.The directory codes are classes of the official classification hierarchy of the *Directory of Community legislation in force*. It contains 20 chapter headings with up to four sub-division levels.

Figure 7 shows an excerpt of a sample document with all information that has not been used removed. The full document can be viewed at http://eur-lex.europa.eu/LexUriServ /LexUriServ.do?uri=CELEX:31991L0250:EN:NOT.

The high number of 3,993 different EUROVOC descriptors were identified in the retrieved documents, each document is associated to 5.37 descriptors on average. In contrast there are only 201 different subject matters appearing in the dataset, with a mean of 2.23 labels per document, and 412 different directory codes, with a label set size of on average 1.29. Note that for the directory codes we used only the assignment to the leaf category as the parent nodes can be deduced from the leaf node assignment. For the document in Figure 7 this would mean a set of labels of $\{17.20\}$ instead of $\{17, 17.20\}$.

5.1. Data Preprocessing

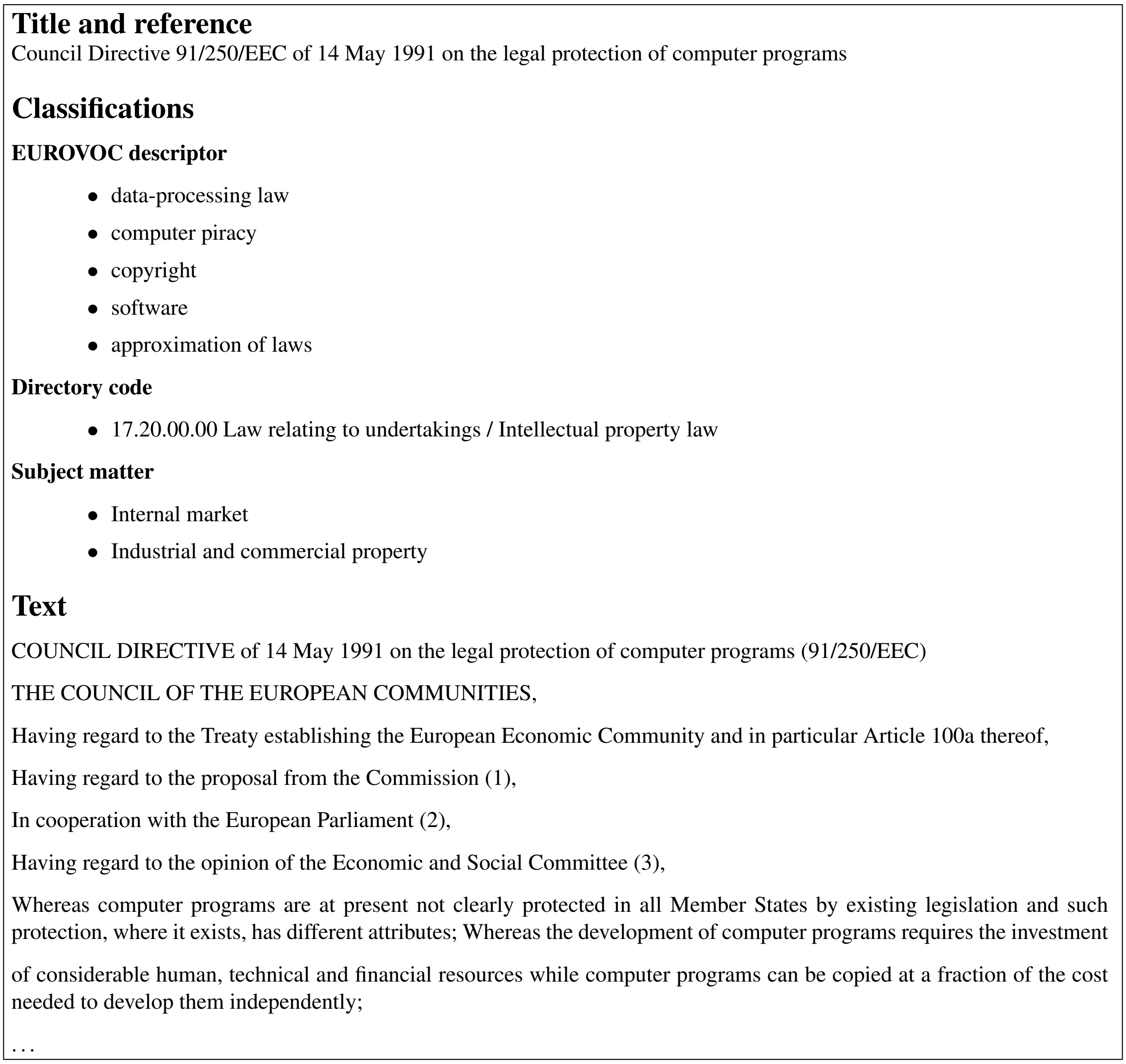
The main text was extracted from the HTML documents, excluding HTML tags, bibliographic notes or other additional information that could distort the results, and was then finally tokenized. The tokens were transformed to lower case, stop words were excluded, and the Porter stemmer algorithm was applied.[6] In order to perform cross validation, the instances were randomly distributed into ten folds. The tokens were projected into the vector space model using the common TF-IDF term weighting (Sebastiani, 2002). In order to reduce the memory requirements, of the approx. 200,000 resulting features we selected the first 5,000 ordered by their document frequency. This feature selection method is very simple and efficient and independent from class assignments, although it performs comparably to more sophisticated methods using chi-square or information gain computation (Yang and Pedersen, 1997). In order to ensure that no information from the test set enters the training phase, the TF-IDF transformation and the feature selection were conducted only on the training sets of the ten cross-validation splits.

[3] http://eur-lex.europa.eu

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

[5] http://europa.eu/eurovoc/

[6] The implementation from the Apache Lucene Project (http: //lucene.apache.org/java/docs/index.html) was used.



Title and reference

Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs

Classifications

EUROVOC descriptor

- data-processing law
- computer piracy
- copyright
- software
- approximation of laws

Directory code

- 17.20.00.00 Law relating to undertakings / Intellectual property law

Subject matter

- Internal market
- Industrial and commercial property

Text

COUNCIL DIRECTIVE of 14 May 1991 on the legal protection of computer programs (91/250/EEC)

THE COUNCIL OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Economic Community and in particular Article 100a thereof,

Having regard to the proposal from the Commission (1),

In cooperation with the European Parliament (2),

Having regard to the opinion of the Economic and Social Committee (3),

Whereas computer programs are at present not clearly protected in all Member States by existing legislation and such protection, where it exists, has different attributes; Whereas the development of computer programs requires the investment

of considerable human, technical and financial resources while computer programs can be copied at a fraction of the cost needed to develop them independently;

...

Figure 7: Excerpt of a EUR-Lex sample document with the CELEX ID 31991L0250. The original document contains more meta-information. We trained our classifiers to predict the EUROVOC descriptors, the directory code and the subject matters based on the text of the document.

6. Evaluation

6.1. Algorithm Setup

For the MMP algorithm we used the IsErr loss function and the uniform penalty function. This setting showed the best results in (Crammer and Singer, 2003) on the RCV1 data set. The perceptrons of the BR and MMP ensembles were initialized with random values. We performed also tests with a multilabel variant of the multinomial Naive Bayes (MLNB) algorithm in order to provide a baseline. For the MLNB we counted the TF-IDF instead of the term frequency values as we obtained improved results by using this additional information about the overall relevance of each term.

6.2. Ranking Performance

The results for the four algorithms and the three different classifications of EUR-Lex are presented in Table 2. The values for IsErr, OneErr, RankLoss and AvgP are shown $\times 100\%$ for better readability, AvgP is also presented in the conventional way (with 100% as the optimal value) and not as a loss function. The number of epochs indicates the number of times that the online-learning algorithms were able to see the training instances. No results are reported for the performance of DMLPP on EUROVOC for more than one epoch.

The first appreciable characteristic is that DMLPP dominates all other algorithms on all three views of the EUR-Lex data, regardless of the number of epochs or losses. For the directory code DMLPP achieve a result in epoch 2 that is still beyond the reach of the other algorithms in epoch 10, except for MMP's IsErr. Especially on the losses that directly evaluate the ranking performance the improvement is quite pronounced and the results are already unreachable after the first epoch. It is also interesting to note that the improvement between epoch 5 and epoch 10 is rather small compared to the previous steps. We can observe this effect also for the MMP algorithm advancing from 10 to 20 epochs (e.g. 40.01 for IsErr and 9.73 % for RankLoss

subject matter	1 epoch				2 epochs			5 epochs			10 epochs		
	MLNB	BR	MMP	DMLPP	BR	MMP	DMLPP	BR	MMP	DMLPP	BR	MMP	DMLPP
IsErr ×100	99.15	61.71	53.61	52.28	57.39	48.83	45.48	52.36	43.21	39.74	49.87	40.89	37.69
OneErr ×100	98.63	30.67	27.95	23.62	26.44	24.4	18.64	22.17	18.82	15.01	20.41	17.08	13.7
RankLoss	8.979	16.36	2.957	1.160	14.82	2.785	0.988	11.40	2.229	0.885	10.29	2.000	0.849
Margin	25.34	59.33	13.04	4.611	54.27	11.94	4.001	44.05	9.567	3.615	40.39	8.636	3.488
AvgP	12.26	62.9	74.71	77.98	66.61	78.05	82.05	71.28	81.71	84.76	73.06	83.19	85.79

directory code	1 epoch				2 epochs			5 epochs			10 epochs		
	MLNB	BR	MMP	DMLPP	BR	MMP	DMLPP	BR	MMP	DMLPP	BR	MMP	DMLPP
IsErr ×100	99.25	52.46	46.08	37.58	45.89	40.78	33.31	40.97	34.27	30.41	37.67	32.52	29.58
OneErr ×100	99.28	44.61	39.42	29.41	37.46	34.27	25.60	32.09	27.62	23.04	28.86	25.83	22.40
RankLoss	7.785	19.30	2.749	1.109	15.12	2.294	0.999	12.10	2.199	0.961	10.17	1.783	0.953
Margin	35.89	96.16	15.47	6.271	77.31	13.30	5.690	62.98	12.30	5.478	53.82	10.20	5.436
AvgP	6.565	57.10	70.00	77.21	63.49	74.61	80.10	68.27	78.83	81.93	71.18	80.28	82.37

EUROVOC	1 epoch				2 epochs		5 epochs	
	MLNB	BR	MMP	DMLPP	BR	MMP	BR	MMP
IsErr ×100	99.58	98.57	98.84	97.92	98.19	97.47	97.23	95.96
OneErr ×100	99.99	48.69	75.89	35.50	41.50	54.41	37.30	40.15
RankLoss	22.88	40.35	3.906	2.779	35.46	4.350	30.96	4.701
Margin	1644.00	3230.68	597.59	433.89	3050.07	694.10	2842.63	761.24
AvgP	1.06	26.90	29.28	46.67	31.58	39.54	35.90	47.27

Table 2: Average losses for the three views on the data and for the different algorithms.

on subject matter for epoch 20, similar behavior for directory code). This partially confirms the results of Crammer and Singer (2003). They observed that after reaching a certain amount of training examples the improvement stops and after that point the performance can even become worse. This point seems to be reached for MMP at the latest at ten epochs on the subject matter and directory code data. It is also interesting to note the behavior on the EUROVOC data as the ranking losses RankLoss and Margin increases from the first epoch on whereas for the other losses it still decreases.

In addition to the fact that the DMLPP outperforms the remaining algorithms, it is still interesting to compare the performances of MMP and BR as they have still the advantage of reduced computational costs and memory requirements in comparison to the (dual) pairwise approach and could therefore be more applicable for very complex data sets such as EUROVOC, which is certainly hard to tackle for DMLPP (cf. Section 6.3.).

For the subject matter and directory code, the results clearly show that the MMP algorithm outperforms the simple one-against-all approach. Especially on the losses that directly evaluate the ranking performance the improvement is quite pronounced. The smallest difference can be observed in terms of OneErr, which evaluates the top class accuracy.

The performance on the EUROVOC descriptor data set confirms the previous results. The differences in RankLoss and Margin are very pronounced. In contrast, in terms of OneErr the MMP algorithm is worse than one-against-all, even after five epochs. It seems that with an increasing amount of classes, the MMP algorithm has more difficulties to push the relevant classes to the top such that the margin is big enough to leave all irrelevant classes below, although the algorithm in general clearly gives the relevant classes a higher score than the one-against-all approach. An explanation could be the dependence between the perceptrons of the MMP. This leads to a natural normalization of the scalar product, while there is no such restriction when trained independently as done in the binary relevance algorithm. As a consequence there could be some perceptrons that produce high maximum scores and thereby often arrive at top positions at the overall ranking. The price to pay for the BR algorithm is a decreased quality of the produced rankings, as the results for RankLoss and Margin are even beaten by Naive Bayes, which is by far the worst algorithm for the other losses.

The fact that in only approximately 4% of the cases a perfect classification is achieved and in only approx. 60% the top class is correctly predicted (MMP) should not lead to an underestimation of the performance of these algorithms. Considering that with almost 4000 possible classes and only 5.3 classes per example the probability of randomly choosing a correct class is less than one percent, namely 0.13%, the performance is indeed substantial.

6.3. Computational Costs

In order to allow a comparison independent from external factors such as logging activities and the run-time environment, we ignored minor operations that have to be performed by all algorithms, such as sorting or internal operations. An overview over the amount of real value addition and multiplication computations is given in Table 6.3. (measured on the first cross validation split, trained for one epoch), together with the CPU-times on an AMD Dual Core Opteron 2000 MHz as additional reference information.

We can observe a clear advantage of the non-pairwise approaches on the subject matter data especially for the prediction phase, however the training costs are in the same order of magnitude. For the directory code the rate for MMP and BR more than doubles in correspondence with the increase in number of classes, while DMLPP profits from the decrease in the average number of classes per instance. It even causes less computations in the training phase than MMP/BR. The reason for this is not only the reduced maximum amount of weights per instance (cf. Section 4.), but particularly the decreased probability that a training example is relevant for a new training example (and consequently that dot products and scores have to be computed) since it is less probable that both class assignments match, i.e. that both examples have the same pair of positive and negative classes. This becomes particularly clear if we observe the number of non-zero weights and actually used weights during training for each new example. The classifier for subject matter has on average 21 weights set per instance out of 443 ($= L(K - L)$) in the worst case (a ratio of 4.47%), and on average 5.1% of them are required when a new training example arrives. For the directory code with a smaller fraction L/K 35.5 weights are stored (3.96%), of which only 1.11% are used when updating. This also explains the relatively small number of operations for training on EUROVOC, since from the 1,802 weights per instance (8.41%), only 0.55% are relevant to a new training instance. In this context, regarding the disturbing ratio between real value operations and CPU-time for training DMLPP on EUROVOC, we believe that this is caused by a suboptimal storage structure and processing of the weights and we are therefore confident that it is possible to reduce the distance to MMP in terms of actual consumed CPU-time by improving the program code.

Note that MMP and BR compute the same amount of dot products, the computational costs only differ in the number of vector additions, i.e. perceptron updates. It is therefore interesting to observe the contrary behavior of both algorithms when the number of classes increases: while the one-against-all algorithm reduces the ratio of updated perceptrons per training example from 1.33% to 0.34% when going from 202 to 3993 classes, the MMP algorithm doubles the rate from 8.53% to 22.22%. For the MMP this behavior is natural: with more classes the error set size increases and consequently the number of updated perceptrons. In contrast BR receives less positive examples per base classifier, the perceptrons quickly adopt the generally good rule to always return a negative score, which leads to only a few binary errors and consequently to little corrective updates. A more extensive comparison of BR and MMP can be found in a previous work (Loza Mencía and Fürnkranz, 2007).

7. Conclusions

In this paper, we evaluated two known approaches for efficiently solving multilabel classification tasks on a large-scale text classification problem taken from the legal domain: the EUR-Lex database. The experimental results confirm that the MMP algorithm, which improves the more commonly used one-against-all or binary relevance approach by employing a concerted training protocol for the classifier ensemble, is very competitive and well applicable in practice for solving large-scale multilabel problems.

subject matter	training	testing
BR	35.8 s 1,675 M op.	8.36 s 192 M op.
MMP	31.35 s 1,789 M op.	6.28 s 192 M op.
DMLPP	326.02 s 6,089 M op.	145.67 s 4,628 M op.

directory code	training	testing
BR	49.01 s 3,410 M op.	11.99 s 394 M op.
MMP	49.63 s 3,579 M op.	11.03 s 394 M op.
DMLPP	313.59 s 2,986 M op.	192.99 s 5,438 M op.

EUROVOC	training	testing
BR	405.42 s 32,975 M op.	56.71 s 3,817 M op.
MMP	503.04 s 40,510 M op.	53.69 s 3,817 M op.
DMLPP	11,479.81 s 17,719 M op.	7,631.86 s 127,912 M op.

Table 3: Computational costs in CPU-time and millions of real value operations (M op.)

The average precision rate for the EUROVOC classification task, a multilabel classification task with 4000 possible labels, approaches 50%. Roughly speaking, this means that the (on average) five relevant labels of a document will (again, on average) appear within the first 10 ranks in the relevancy ranking of the 4,000 labels. This is a very encouraging result for a possible automated or semi-automated real-world application for categorizing EU legal documents into EUROVOC categories.

In addition we presented an algorithm that reformulates the pairwise decomposition approach to a dual form so that it is capable to handle very complex problems and therefore to compete with the approaches which use one classifier per class. It was demonstrated that decomposing the initial problem into smaller problems for each pair of classes achieves higher prediction accuracy on the EUR-Lex data, since DMLPP substantially outperformed all other algorithms. This confirms previous results of the non-dual variant on the large Reuters Corpus Volume 1 (Loza Mencía and Fürnkranz, 2008). The dual form representation allows for handling a much higher number of classes than the explicit representation, albeit with an increased dependence on the training set size. We are currently investigating variants to further reduce the computational complexity. Despite the improved ability to handle large problems, DMLPP is still less efficient than MMP, especially for the EUROVOC data with 4000 classes. However, in our opinion the results show that DMLPP is still competitive for solving large-scale problems in practice, especially considering the trade-off between runtime and prediction performance.

For future research, on the one hand we see space for improvement for the MMP and pairwise approach for instance by using a calibrated ranking approach (Brinker et al., 2006). The basic idea of this algorithm is to introduce an artificial label which, for each example, separates the relevant from irrelevant labels in order to return a set of classes instead of only a ranking. On the other hand, we see possible improvements by exploiting advancements in the perceptron algorithm and in the pairwise binarization, e.g. by using one of the several variants of the perceptron algorithm that, similar to SVMs, try to maximize the margin of the separating hyperplane in order to produce more accurate models (Crammer et al., 2006; Khardon and Wachman, 2007), or by employing a voting technique that takes the prediction weights into account such as the weighted voting technique by Price et al. (1995).

Acknowledgements

This work was supported by the EC 6th framework project ALIS (Automated Legal Information System).

8. References

Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

Klaus Brinker, Johannes Fürnkranz, and Eyke Hüllermeier. A Unified Model for Multilabel Classification and Ranking. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI-06)*, 2006.

Koby Crammer and Yoram Singer. A Family of Additive Online Algorithms for Category Ranking. *Journal of Machine Learning Research*, 3(6):1025–1058, 2003.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

Yoav Freund and Robert E. Schapire. Large Margin Classification using the Perceptron Algorithm. *Machine Learning*, 37(3): 277–296, 1999.

Johannes Fürnkranz. Round Robin Classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

Chih-Wei Hsu and Chih-Jen Lin. A Comparison of Methods for Multi-class Support Vector Machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

Roni Khardon and Gabriel Wachman. Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research*, 8:227–248, 2007.

Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Handwritten Digit Recognition by Neural Networks with Single-Layer Training. *IEEE Transactions on Neural Networks*, 3(6):962–968, 1992.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Eneldo Loza Mencía and Johannes Fürnkranz. An evaluation of efficient multilabel classification algorithms for large-scale problems in the legal domain. In *LWA 2007: Lernen - Wissen - Adaption, Workshop Proceedings*, pages 126–132, 2007.

Eneldo Loza Mencía and Johannes Fürnkranz. Pairwise learning of multilabel classifications with perceptrons. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (to appear)*, 2008.

David Price, Stefan Knerr, Leon Personnaz, and Gerard Dreyfus. Pairwise Neural Network Classifiers with Probabilistic Outputs. In *Advances in Neural Information Processing Systems*, volume 7, pages 1109–1116. The MIT Press, 1995.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

Shai Shalev-Shwartz and Yoram Singer. A New Perspective on an Old Perceptron Algorithm. In *Learning Theory, 18th Annual Conference on Learning Theory (COLT 2005)*, pages 264–278. Springer, 2005.

Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-486-3.

Judicial Precedents Processing Project for Supporting Japanese Lay Judge System

Rafal Rzepka[*], Hideyuki Shibuki[†], Yasutomo Kimura[‡], Keiichi Takamaru[§], Masafumi Matsuhara[¶], Koji Murakami[‖]

Abstract

Abstract. In this paper we introduce our ideas for a system which would help lay judges to prepare themselves for trials without reading vast amounts of legal texts regarding past cases. After introducing new Japanese "lay judge system" and its possible problems, we describe an outline for support tool using NLP techniques. By using topic-tailored named entity recognition, coreference resolution, template element extraction and scenario template creation, we aim at building a helpful tool with visualizing function in order to help law amateurs not only to estimate the penalty for a given case under particular set of conditions, but also to learn about how those conditions influence verdicts of professionals.

Keywords: NLP, legal information retrieval and extraction, legal text processing

1 Introduction

In May 2009 "lay judge system"[1] (sometimes called "citizen judge system" or "lay assessor system"[2]) will be introduced in Japan. This revision to the Criminal Procedure Code, in which private citizens join with professional judges in trying serious criminal cases, stipulates that any eligible voter aged 20 or older can be selected by lottery to serve as a lay judge in hearings on serious crimes including homicide, and states that participation in such hearings is a public obligation. Lay judges will help decide whether a defendant is guilty and, in the event of a guilty verdict, determine the sentence to be handed down (which is different from usual jury systems in other countries). Our goal is to construct a NLP-based tool that could support a lay judge helping him or her to acquire knowledge on judicial precedents of similar cases. A person without any law experience chosen for a lay judge will need a lot of time to prepare for a trial, therefore our tool could save his or her time. We also hope it could be useful for anyone interested in justice administration, history or tendencies of crime and punishment.

Preparations for switching the current system of professional judges to lay judges are underway but judicial circles are rather groping their way in the dark as this is a new concept in Japan (though not unprecedented, jury system existed in Japan before World War II). One of the tangible results of above mentioned preparations is digitalizing documents containing judicial precedents and making them available online. However, for somebody who is not a lawyer, it is a very difficult task to read and entirely understand very big number of documents full of specific legal terms and

[*] Graduate School of Information Science and Technology, Hokkaido University, Kita-ku Kita 14 Nishi 9, Sapporo, 060-0814, Japan

[†] High-Tech Research Center, Hokkai-Gakuen University, Minami 26, Nishi 11-1-1, Chuo-ku, Sapporo, 064-0926, Japan

[‡] Department of Information and Management Science, Otaru University of Commerce, 3-5-21, Midori, Otaru, 047-8501, Japan

[§] Utusnomiya Kyowa University, 1-3-18, Odori, Utsunomiya, 320-0811, Japan

[¶] Department of Software and Information Science, Iwate Prefectural University, Iwate, 020-0193, Japan

[‖] Center for Innovation Systems Research, Tokyo Institute of Technology, 4259-S1, Nagatsuda, Midori-ku, Yokohama, 226-8303, Japan

Table 1: Comparison of Usual Jury Systems and Japanese *saiban'in* Lay Judge System

	Jury Systems	Lay Judge Systems	Japanese Lay Judge System
Pro Judges Participation	only jurors	together with jurors	together with jurors
Decision of guilt	yes	yes	yes
Decision of sentence	no	yes	yes
Lenghth of assignment	for a case	tenure	for a case
Jury selection	at random	recommendation	at random

expressions. Because this is not helpful enough and no support-tool exists, we decided to apply Natural Language Processing techniques to aid non-experts in retrieving desired information and interpreting it correctly. The first step of our project to build a lay judge support tool is to prepare algorithms for recognizing named entities for assessment of culpability and methods for tagging topology attributes. In this paper we first explain basic characteristics and possible problems of a new lay judge system, then introduce outline of the proposed system.

2 Purpose

2.1 Japanese Lay Judge System

Usually there are two different types of juridical systems where citizens play active role in a trial - jury system and lay judge system. The lay judge system is adopted in Germany, France, and Italy, while in United Kingdom and the United States, the jury system is in force. The main differences between jury system, lay judge system and Japanese lay judge system are summarized in Table 1. Below we underline those differences in a more specific manner.

In the jury system, citizens inquire into a case independently of the professional judges, but in *saiban'in* system jurors and professional judges consult with each other on an equal basis. Moreover, jurors in jury system basically only authorize the fact of crime and the judge decides on sentence according to the law which fits the case best in his opinion, while jurors in the Japanese lay judge system will not only authorize the fact of crime but also decide sentencing (professional judges are supposed to determine suitable law only). In a lay judge system a judge and jurors constitute one council unit which authorizes the fact of crime and decides the suitable regulation. A lay judge is elected for a tenure because high degree of professionalism is requested, while a Japanese lay judge will be selected at random for each case. As the same level of professionalism is needed, short-term duty will bring a danger of having too little experience for a lay judge to determine appropriate sentencing.

Another problem is that until now there were only professional judges in Japan and citizen participation in trials practically didn't exist. All the information on crime and punishment has been passing to the society mostly through mass media. Principally Japanese law is based on "no punishment when questionable" policy, although Japanese media tend to show suspects as a culprits. It is obvious that court must be independent and unbiased, but in many cases it is difficult to sentence without being influenced by outside factors. For such reasons the use of past cases documents on judicial precedents is highly recommended to help keeping objectivity of judgment and easy access to such data should be supported by NLP techniques tailored for field of legal texts.

Even if a lay judge is not influenced by mass media and does not lose his or her objectivity, it still will be a problem to decide on the sentence for a person without legal experience. In Japan penal regulations regarding punishment of given type of crime are stated in Criminal Code, though there are many factors which influence the final sentence. For example in case of homicide, there is a wide range of possible verdicts from 5 years of imprisonment, through life imprisonment to

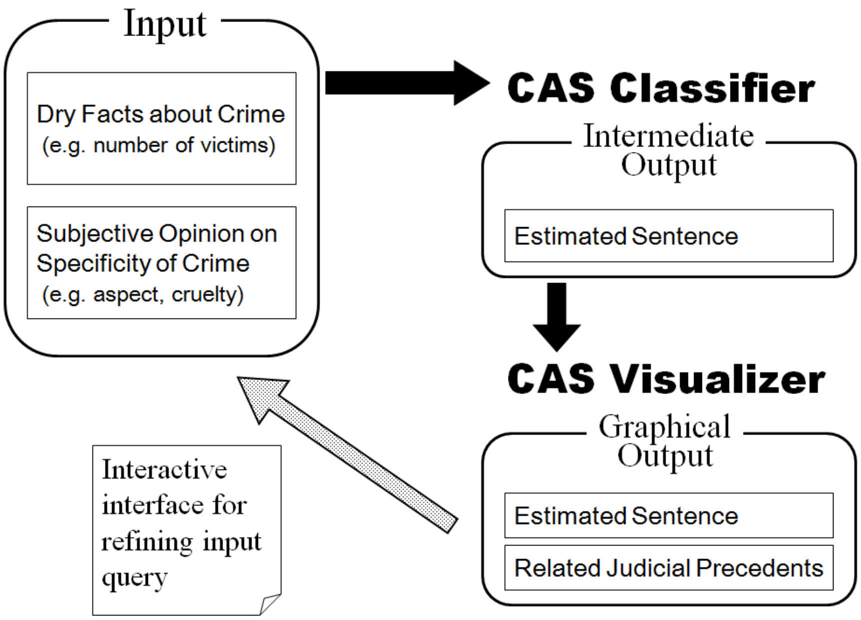


Figure 1: Outline of our support tool for lay judges

capital punishment. Even if the commitment of murder is an undeniable fact, circumstances can influence the sentence (for instance drastically changing the length of imprisonment), however such calculation is a very difficult task without knowledge on similar cases.

A duty of applying a correct law to given case belongs to a professional judge but a lay judge still has to realize all the specificity of a crime being analyzed - it may be a homicide, an involuntary manslaughter or manslaughter by negligence. In Japan it is clearly stated what punishment comes with which type of crime - death sentence for homicide, more than three years for involuntary manslaughter, less than five hundred thousands yen for manslaughter by negligence, etc. Difference between homicide and involuntary manslaughter lies in whether somebody had murderous intent or not, and difference between involuntary manslaughter and manslaughter by negligence lies in whether somebody had intention to hurt other person or not. Therefore, if one person hits another person in order to cause injury but had no intention to take life, life imprisonment or death penalty are not apprioprate. However, an amateur can easily neglect objective analysis, especially under the influence of emotions, and forget about taking all the aspects into consideration. A tool which would make a lay judge analyze the case by comparing it to similar cases by inputting all available factors is desirable and this is the main reason we want to propose system extracting features of culpability assessment from precedents to support lay judges in their sentencing process. In this paper we abbreviate this process to CAS (Culpability Assessment for Sentencing).

3 Our Vision of CAS Processing

3.1 Outline

The idea of system that we have started to build is shown in Figure 1. After inputting a summary of given case, CAS Classifier estimates a sentence best fitting the inputted conditions. For easier understanding of the conditions' influence on sentencing, the output is visualized by CAS Visualizer. The summary includes not only dry facts as number of victims or method of murder, it also lets to input subjective observations as presence of maliciousness, planning, level of impact on the society or defendant's attitude. Subjective observations differ from dry facts as there is no perfect agreement on personal views which can change during the trial. Because those feelings towards a

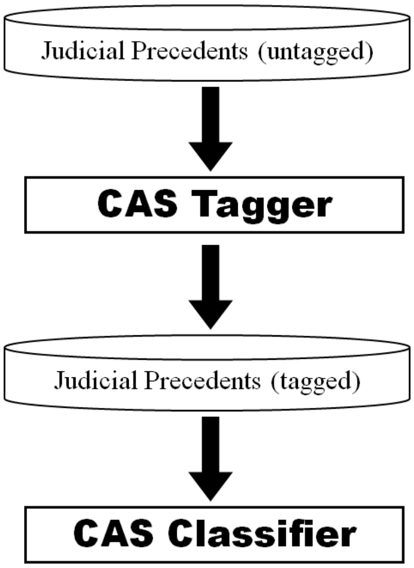


Figure 2: Learning process of CAS Classifier

case can influence the sentencing process (to give a simple example - the more malicious the crime seems to be, the heavier punishment can be chosen), there is a need for representing the change of those influences in a easily understandable manner. We plan to make a system that not only shows the example sentencing possibilities for a set of condition, but also vice-versa - what kind of conditions are common for a given type of penalties, for example the heaviest ones as death penalty or life imprisonment. To achieve this goal, we will implement CAS Visualizer which is planned to help a lay judge not only to summarize previous cases for a set of condition, but also to interactively learn about sentencing tendencies.

The mentioned before CAS Classifier needs to know the correspondence between the sentencing examples and inputted conditions in order to calculate the most possible sentence. Because manual creation of such set of mutual correspondence is close to impossible, they must be learned automatically from the precedents. To prepare the CAS Classifier, we use the digitalized texts on previous verdicts as a learning data set. Those texts are available for anyone online and they describe details on reasoning used for sentencing.

Learning process of CAS Classifier is shown in Figure 2. Usually one needs a big amount of tagged documents for achieving high accuracy of classification, but there is only a few electronic resources for trial records. Furthermore, manual tagging is a very laborious and costly undertaking, therefore we decided to create a CAS Tagger which is going to automatically tag the texts by using smaller set of manually tagged texts. In this section we will describe more details on and particular steps of creating CAS Classifier, CAS Tagger and CAS Visualizer - essential elements of our concept.

3.2 CAS Classifier

In CAS Classifier, assumption of becoming the same sentencing can be made if there was the same judicial precedent. It is of course impossible that exactly the same case exists, though for sentencing estimation there is a need to narrow the set of judicial precedents. It is also necessary to set appropriate condition tag to a sentence given for a particular case. For instance, it seems that there is no influence on the increase and decrease of penalty if a murder had place in Tokyo or in Osaka. Oppositely, when the degree of defendant's regret is quite high then possibility for lighter penalty appears, while a defendant who doesn't seem to regret his own crime will probably punished with heavier penalty. Thus, it can be said that defendant's attitude tag is necessary but tag for place where the event occured is not – regarding the influence on sentencing. Therefore, the first task is to decide which conditions should be tagged to reflect on the penalty judgment.

To solve this task, the standards used in an actual judicial decision become reference. For instance, one of the most famous standards are Nagayama Criteria which can be called death sentencing guidelines for Japanese courts. The Criteria became a standard after a trial of Norio Nagayama who was sentenced to death for killing four people in a shooting spree in 1968 at age 19. It consists of following nine items:

- character of crime
- motive of crime
- crime method, with special emphasis on presence of obstinacy and cruelty
- importance of crime result, with special emphasis on number of murdered victims
- bereaved family's feelings
- influence on society
- age of criminal
- previous offense (criminal record)
- circumstances after the crime

Such standards are suitable for human judgment, however it is necessary to arrange them into a set of expressions suitable for being processed by the computer. Moreover, not always all circumstances can be inputted for a given case, therefore the necessity of analyzing judicial precedents appears.

Next task is to solve problem concerning the values of the condition tags. For instance, in Japanese law, the difference between 27 years old and 28 years old does not influence verdict much, while the difference between 17 years old and 18 years old criminals has a big impact on sentencing standards. For that reason it is better to perform tagging for machine learning not by setting numerical tag for AGE entity, but to create more informative values as JUVENILE and ADULT to achieve higher accuracy of classification. Because types of tags differ in a dimension of influencing verdicts, there is a need to set tag values based on what influence the tagged factors had on sentence in an actual judicial precedent. Moreover, the paraphrase processing of tagged expressions will be probably needed to assure appropriate input data for CAS Classifier. Such processing can be considered as a kind of classification, and if we can assure the input expression of satisfactory level of paraphrasing accuracy, then machine learning should become applicable.

The last task is to solve problems concerning the classification output. We are aiming at improvement of the classification accuracy by giving data with a large amount of tags made by CAS Tagger for an actual judicial precedents as training data, but there is still high possibility of cases with low similarity to the past trials. CAS is supposed to be a support system for lay judges and the final decision is assumed to be entrusted by a human. Therefore, it might be more profitable to show the index of penalties seen from different angles of aspect than to narrow the range of possible penalties for a particular case. Thus it is better not only to set appriopriate level of similarity, but also to develop a function of bringing together features common for a given viewpoint (aspect of circumstances). For achieving this function we need to construct a typology of, for example, semantic relations between condition tags and to perform bi-directional relating of conditions and penalties.

3.3 CAS Tagger

The purpose of CAS Tagger is to generate a training set for CAS Classifier. This processing can be considered as a kind of information extraction from the raw texts of judicial precedents where

CAS Classifier's input are sentenced penalty output and suitable feature value. In general, information extraction task should consist of following subtasks: named entity recognition, coreference resolution, template element extraction and scenario template creation. In our case the usage is limited to verdicts of judicial precedents, therefore it is necessary to remodel them to fit the needs of sentencing support.

For named entity recognition, appropriate tag must be given to the judicial precedent text as a feature of CAS Classifier. For instance, we can improve the overall efficiency by creating directly a CRIMINAL'S AGE tag instead of simple AGE tag. First of all, it It would be the best for the efficiency to set tags identical to ones of CAS Classifier. Next, although the noun phrases are often objects of usual named entity recognition process, it is necessary to target description expressions such as "malignancy" or "cruelty" which are important in the viewpoint of sentencing judgment. The number of rules is expected to increase; therefore it is necessary to prepare the training data for named entity recognition as machine learning will be needed.

Finally, we plan to limit the range of objects to gain better accuracy by using the judicial precedent texts structure analysis made by [3] and [6]. It is also possible to provide restrictions based on rules taken directly from Japanese law as, for example, one saying that upper bound for a sentence of imprisonment for a definite term is 30 years. Applying such knowledge in this stage would most likely increase the overall performance of our proposed system. We expect that in case of coreference resolution, as unique naming is being used for particular case of judicial precedent, it will be much easier than in case of usual documents without specific character and technical terms.

We believe that the templates for template element extraction can be appropriately created by hand as we limit the system to penalty estimation. Moreover, it is most likely that the processing load of element extraction decreases if appropriate tag can be given. However, because vagueness always exists, it is necessary to ignore cases where two or more crimes took place. One act of crime is usually described in one paragraph, therefore the same techniques as in named entity recognition process will become important.

While creating scenario templates it is certain that the final shape of a template will be the input format of CAS Classifier. Because it is necessary to deal with concurrent offenses, two or more crime templates should be able to be merged. Moreover, while presenting one's penalty decision, it is crucial to understand the relation between dry facts and subjective judgments which will become a material for further objective judgment – for that reason we might need processing that extracts such relations.

We also assume that there is a possibility that the output results of CAS Tagger can be applied besides CAS Classifier, and used as preprocessing for various applications working on legal texts, not only for culpability assessment purposes.

3.4 CAS Visualizer

The concept of CAS Visualizer is to present the following information with the expressive form which should be easily understood by a law amateur.

- What is the estimation of penalty range under present circumstances?
- What is the penalty that seems to be appropriate?
- Which judicial precedents are similar?
- How do they resemble the current case?
- Which circumstance should become an issue?
- How the change of circumstances influence the sentencing?

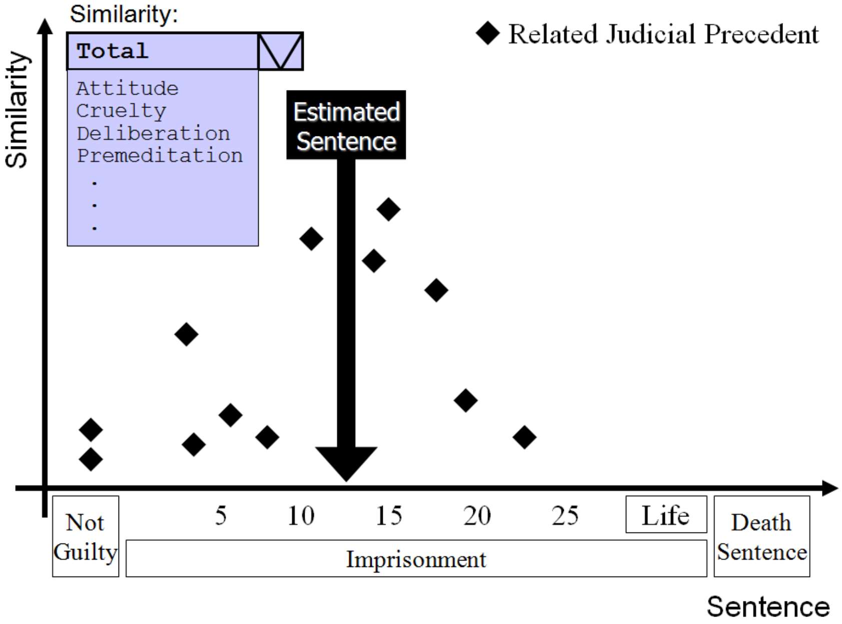


Figure 3: Image of CAS Visualizer - choosing the subjectivity factor

Figures 3 and 4 show the image of CAS Visualizer. Penalty level from innocence through years of imprisonment, life imprisonment and the capital punishment is shown on a horizontal axis, while the degree of similarity with past cases is on vertical axis. Similar judicial precedents are expressed as scatterplots. Penalty estimated "accurate" by CAS Classifier becomes an indicator for a lay judge. In Figure 3 there is a dropdown list where a lay judge can change a viewpoint and see how penalty is changing according to a suspect behavior during or after committing crime (for example if there was cruelty, if there is regret afterwards, etc). It becomes easy to understand the influences if we group the judicial precedents inside the selected viewpoint into high similarity clusters displayed in different colors. A judge can click (as shown in Figure 4)a point that in his opinion is most accurate and see the differences between the conditions. By visualizing process, issues should become clear and a judge can interactively narrow the range of possible penalties for given set of conditions.

3.5 Related Work

Quite a number of papers regarding NLP for legal texts was written in Japan and they also used judicial precedents. Two expert systems intended for the law were widely noticed – legal reasoning systems constructed in 1980's trying to solve legal problems usually faced by law professionals[14][15]. When compared to usual natural language texts, legal texts have many restrictions on terms and grammar usage, rare phrases and grammatical expressions often appear[8][9]. By using those restrictions, analysis of syntax structure and semantic analysis based on case frames were performed[10] [11] [12] [13]. On the other hand, many researchers of the judicial precedents have aimed to retrieve a similar cases data by using the information retrieval technologies. In case of searching records on past trials it is known that the search accuracy doesn't improve only by using the vector space model utilizing the weight of TF*IDF, therefore the retrieval methods using structure of the judicial precedents were proposed[5][6] [7]. Above mentioned researches do not consider penalty level because they were developed as systems for information retrieval only. Currently Muramatsu et al. are trying to develop a tool that displays the logical structure of the sentence (verdict) and guesses the judge's decision upon previous cases[4]. As they also use tags, their work can be seen as the closest one to ours, however their tagging stays within margins of reason, premise and court deci-

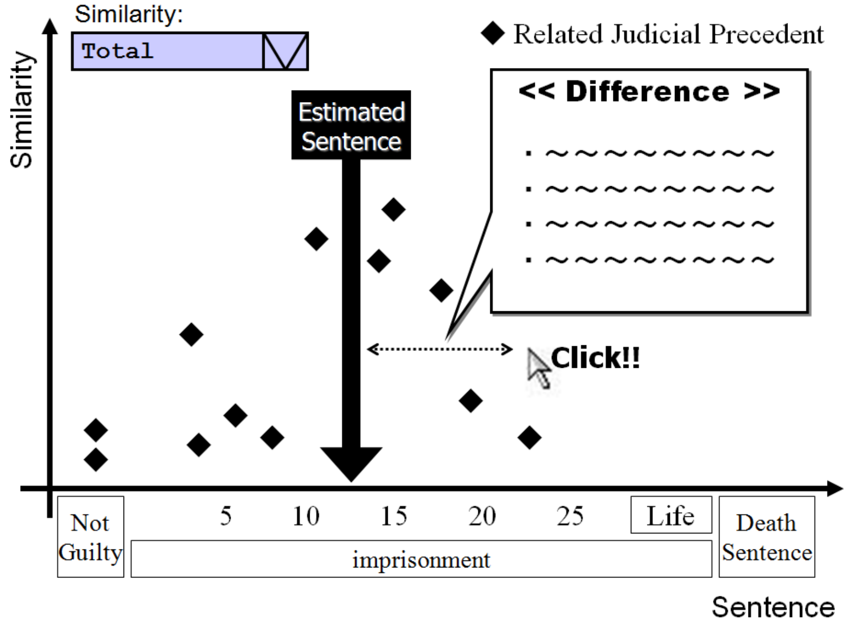


Figure 4: Image of CAS Vizualizer - checking the differences

sion. In our case, we perform estimation of penalty fitting to conditions decided by CAS Classifier and automatic tagging needed for estimation is performed.

4 Conclusion

In our paper we introduced our idea of a supporting tool for law amateurs who will be selected randomly in Japan from May 2009 to become a ”lay judge”, a specific type of juror. We described possible problems of the new system and proposed a set of NLP techniques which should increase the tool’s accuracy by tailoring them to work on very specific data of legal text on judicial precedent. There is still a lot of ambiguity about the new juridical system in Japan and people are worried if they can handle such a difficult task as becoming a ”lay judge”. Our proposed system would not only help to estimate possible sentence but also to learn the mechanisms of decision making and influences of conditions. We are technically ready to start our project but first we decided to conduct a survey and find a way to discuss the topic with legal text processing specialists. Before the tool development process, we also plan to consult our ideas (and cooperate if possible) with professional lawyers in order to use their opinions and advices to build a tool as useful as possible.

References

[1] “Start of the Saiban-in System” (Pamphlet by the Ministry of Justice), http://www.moj.go.jp/SAIBANIN/pdf/pamphlet-e.pdf

[2] English translation of the Saiban’in Law by Anderson and Dear http://www.hawaii.edu/aplpj/pdfs/v6.01_Anderson.pdf

[3] H. Hiroki, Y. Yasumura, D. Katagami, K. Nitta, “The trial of the judicial precedent document retrieval focusing on the structure”, The 18th Annual Conference of the Japanese Society for Artificial Intelligence, 2004. (in Japanese)

[4] M. Muramatsu, Y. Yasumura and K. Nitta, “A Tagging Tool for Logical Structure of Legal Sentences”, Technical Report of IEICE, KBSE2001-49, pp.1-7, 2002. (in Japanese)

[5] M. Harada and R. Suzuki, “Judicial CAse REtriever JCare from Japanese Query Sentences based on Semantic Graph Matching”, IPSJ SIG Notes, 2001-FI-61-3, Vol.2001, No.20, pp.15-22, 2001. (in Japanese)

[6] H. Egoshi, D Katagami and K. Nitta, “Judicial Precedent Retrieval Using the Structure of Documents”, IPSJ SIG Notes, Natural language understanding and models of communication, Vol.2005; No.11 (DD-48), pp.1-8, 2005. (in Japanese)

[7] M. Ishikawa and K. Nitta, “Toward the improvement of precedent retrieval on Legal Reasoning System HELIC-II,” IPSJ SIG Notes, FI, Vol.96, No.88, pp.97-102, 1996. (in Japanese)

[8] K. NAGANO, H.Nagai, T. Nakamura and H. NOMURA, “A Restricted Linguistic Model Based on Verb Functions for Law Sentences”, IPSJ SIG Notes, Natural language understanding and models of communication, Vol.93, No.51, pp.25-32, 1993. (in Japanese)

[9] H. Iwamoto, K. Nagano, H. Nagai, T. Nakamura and H.Nomura, “An Analysis of coordinate Structures and its Application to A controlled Linguistic Model for Law Sentences”, IPSJ SIG Notes, Natural language, Vol.93, No.101, pp.17-24. (in Japanese)

[10] K. Tanaka, I.Kawazoe and H. Narita, “Standard Structure of Legal Provisions”, IPSJ SIG Notes, Natural language, Vol.93, No.79, pp.79-86 1993. (in Japanese)

[11] K. Tanaka, “About Semantic Function of the Legal-Effect’s Restrictive Part”, IPSJ SIG Notes, Natural language, Vol.98, No.21, pp.1-8, 1998. (in Japanese)

[12] K.Hiramatsu, H.Nagai, T.Nakamura and H.Nomura, “A Controlled Linguistic Model for Legal Sentences Based on Legal Condition-Effect Structure”, IPSJ SIG Notes, Natural language, Vol.96, No.87, pp.21-27, 1996. (in Japanese)

[13] K. Kaneiwa and S. Tojo, “A Legal Reasoning System with Event and Property Interpretation for Legal Knowledge”, Transactions of Information Processing Society of Japan, Vol.40, No.7, pp.2892-2904, 1999. (in Japanese)

[14] H. Yoshino, “Legal Expert System as Legal Reasoning System”, IPSJ SIG Notes, Vol.86, No.45, 1986.(in Japanese)

[15] K. Nagano, H. Iwamoto, H. Nagai and H. Nomura, “Analysis of Sentence Endings and its Application to Linguistic Model for Law Sentences”, , IPSJ SIG Notes, Natural language, Vol.1992, No.33, pp75-82, 1992. (in Japanese)

Evaluation Metrics for Consistent Translation of Japanese Legal Sentences

OGAWA, Yasuhiro   IMAI, Kazuhiro   TOYAMA, Katsuhiko

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN
yasuhiro@is.nagoya-u.ac.jp

Abstract

We propose new translation evaluation metrics for legal sentences. Since most previous metrics have been proposed to evaluate machine translation systems, they prepare human reference translations and assume that several correct translations exist for one source sentence. However, the existence of several translations of one legal expression confuses readers who might mistakenly believe that different translations indicate different meanings. Therefore, since translation variety is unacceptable and consistency is crucial in legal translation, we propose two metrics to evaluate the consistency of legal translations and illustrate their performances by comparing them with other metrics.

1. Introduction

Recently, the social demand for the translation of Japanese statutes into foreign languages has been increasing to conduct international transactions more smoothly, to promote international investment in Japan, to support legal reform in developing countries, and so on. Since Japanese statutes have been individually translated by government ministries or private publishing companies, translation equivalents have been inconsistent among translated documents. For example, the legal term “弁護士 (*bengoshi*)” has been translated as “*attorney*,” “*barrister*,” and “*lawyer*,” which have different meanings in English. Therefore if “*attorney*” is used in one document, while “*lawyer*” is used in another document for the same term “弁護士,” it is hard to recognize that both English words denote the same word in the source documents, or, in some cases, they may confuse readers. For this reason, the same translation equivalent should be used for the same term: consistent translation is required.

To solve this problem, the Japanese government has compiled a *Japanese-English Standard Bilingual Dictionary*[1] (Study Council, 2006) (Toyama et al., 2006) for legal technical terms occurring in Japanese statutes, which includes about 3,400 Japanese entries and about 4,200 English equivalents. Now, Japanese statutes are being translated in compliance with this dictionary by the government. Then, the next task is quality evaluation of the translations, which should be done in compliance with the dictionary.

However, since one term sometimes has several translation equivalents, a suitable one in context should be used in a translation. For example, in the Standard Bilingual Dictionary, the term “免除する (*menjo-suru*)” has six equivalents: “*release*,” “*exempt*,” “*waive*,” “*exculpate*,” “*remit*,” and “*immunize*.” We should choose the most suitable one among them depending on the context. Although notices for the choice might be roughly given to some equivalents in the dictionary, registering every detailed criterion for the choice in the dictionary is difficult. Thus it is insufficient to only rely on the dictionary for consistent translations.

Therefore we need an automatic evaluation metric for consistent translations. Several translation metrics have been proposed: BLEU (Papineni et al., 2002), Word Error Rate (WER), Position independent Word Error Rate (PER) (Leusch et al., 2003), METEOR (Banerjee and Lavie, 2005) and NIST (Doddington, 2002). However, these metrics were designed to evaluate machine translation systems and do not evaluate human translations. Since their basic ideas are to compare machine translations with human reference translations that are considered correct, they require such reference translations.

On the other hand, we cannot prepare human reference translations for the evaluation of legal translations. In fact, if we can prepare *correct* reference translations, we no longer need other translations. Therefore, we must prepare alternative references.

In this study, we focus on the fact that most Japanese legal sentences are described in terms of fixed expressions. This is because the Cabinet Legislation Bureau reviews most Japanese statutes and controls the use of legal terms and expressions in the statutes during the process of drafting. From the viewpoint of consistency, the same fixed expression should have the same translation. For this purpose, we used a legal parallel

[1] http://www.kl.i.is.nagoya-u.ac.jp/told/

corpus and compared translations with it.

In particular, for the corpus, we used the translations of Japanese statutes released by the Japanese government[1] (Study Council, 2006), including 17,793 Japanese sentences. We use this parallel corpus instead of human reference translations. However, the translations in the corpus may not be translations of source sentences but translations of similar sentences to the sources. We call such translations *pseudo reference translations* (PRTs).

We tried to use the BLEU metric (Papineni et al., 2002) to compare a candidate translation with PRTs. However, the BLEU metric is not convenient for our comparisons, since PRTs are not translations of source sentences. To solve this problem, we modified the BLEU metric tailored with PRTs and we named it **CIEL**.

We applied the CIEL metric to three different translations of the Labor Standard Act: by the Japanese government, a publishing company, and the Google translation tool. The CIEL metric distinguished the translations by the government from those by the publishing company, but the BLEU metric could not. In addition, we compared the CIEL metric with other evaluation metrics and showed its effectiveness.

This paper is organized as follows: in Section 2, we introduce the BLEU metric as the baseline. Next, we propose our evaluation metrics in Section 3. Then we describe some evaluation experiments in Section 4. Finally, Section 5 is a conclusion.

2. BLEU

The BLEU metric (Papineni et al., 2002) is an automatic evaluation metric for machine translation. Its basic idea compares *n*-grams occurring in the candidate translation, which is a machine translation sentence for a given source sentence, with *n*-grams occurring in the human reference translations. Since several translations are possible for one source sentence, the BLEU metric prepares multiple human translations as references. For comparison, the following precision score p_n is calculated:

$$p_n = \frac{\sum\limits_{S \in Candidates} \sum\limits_{n\text{-}gram \in S} Count_{clip}(n\text{-}gram)}{\sum\limits_{S \in Candidates} \sum\limits_{n\text{-}gram \in S} Count(n\text{-}gram)}, \quad (1)$$

where $Count(n\text{-}gram)$ is the number of occurrences of *n-gram* in the candidate translation *S*. $Count_{clip}(n\text{-}gram)$ is also the number of occurrence of *n-gram* in *S*, but if it is greater than the maximum number of occurrences of *n-gram* that occurs in any single reference translation, $Count_{clip}(n\text{-}gram)$ is equal to the maximum number. Notice that if *n-gram* does not occur in any reference translations, $Count_{clip}(n\text{-}gram)$ is 0.

Next, if the candidate translation is shorter than its reference translations, the denominator of the above formula becomes smaller so that p_n becomes larger. To penalize this situation, the BLEU metric computes brevity penalty (BP):

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}, \quad (2)$$

where c is the length of the candidate translation and r is its effective reference length.

Finally, introducing positive weights w_n based on the value of n, the final BLEU score is defined as follows:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \quad (3)$$

Usually, the upper of n is set to be $N = 4$ and uniform weights $w_n = 1/N$. Using from unigrams to 4-grams, the BLEU metric evaluates both adequacy and fluency of candidate translations, where the adequacy indicates how much information is retained in the translation and the fluency indicates to what extent the translation reads like good English.

3. Evaluation Metric Considering Consistency

The BLEU metric needs several human reference translations that are considered *correct*. For the evaluation of translation of legal sentences, however, we cannot prepare reference translations. In fact, if we have a *correct* translation of a legal sentence, we do not need to evaluate other translations. So we have to evaluate human translations without *correct* references.

Here, notice that Japanese legal sentences have many fixed expressions. For example, the sentences that provide effective dates of each act have the following expressions:

Source 1. この法律は、会社法 の施行の日から施行する。

Source 2. この法律は、行政手続法 の施行の日から施行する。

For consistent translation, such fixed expressions as "この法律は、... の施行の日から施行する。," shown as underlined, should be translated into identical expressions. Therefore Sources 1 and 2 should be respectively translated as follows:

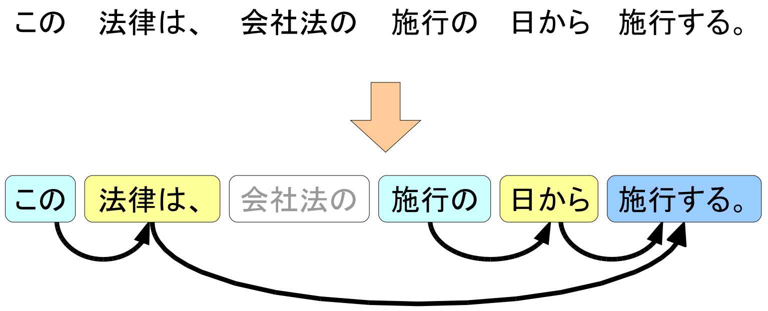


Figure 1: *Bunsetsu* deletion.

Translation 1. This Act shall come into force as from the date of enforcement of the Companies Act.

Translation 2. This Act shall come into force as from the date of enforcement of the Administrative Procedure Act.

We used a parallel corpus of Japanese statutes to evaluate such consistency. First, we retrieved similar sentences to a given source sentence and collect their translations, which can be considered reference translations. However, since such translations may not be translations of source sentences, we call them *pseudo reference translations* (PRTs). To compile PRTs, we used legal translations released by the Japanese government[1] (Study Council, 2006). We assume they are suitable translations since they were made in compliance with the Standard Bilingual Dictionary to improve consistency and reliability of the translations. Thus, the translations in the PRTs can be considered adequate and fluent in terms of consistency.

We describe the details of the compilation of PRTs in Section 3.1. and how to evaluate the translations in Sections 3.2. and 3.3., respectively.

3.1. Acquisition of Pseudo Reference Translations

We used a hierarchical clustering method (Jain et al., 1999) to obtain a set of PRTs. We divided the source sentences in the corpus into clusters and selected the closest one to a given source sentence. Since such clusters contain similar sentences to the source, we collected their translations as PRTs. The following shows the details of the clustering method.

First, we split a set of source sentences since the cost of clustering tasks for all sentences is considered to be too high. Since main predicates, which play an important role in sentences, occur at the end of sentences in Japanese, we split them by their last morphemes.

Next, we deleted all *bunsetsus*[2], except the last one, those depending on the last one, and those depending on them. This is to delete non-fixed expressions from the sentences.

For example, the following *bunsetsus* are left after the deletion in Source 1 consisting of six *bunsetsus*:

1. “施行する (*shall come into force*)”; this is the last *bunsetsu*.
2. “法律は (*Act*),” “日から (*as from the date*)”; these depend on the last *bunsetsu* “施行する.”
3. “この (*This*),” “施行の (*of enforcement*)”; “この” depends on “法律は” and “施行の” depends on “日から.”

This result is illustrated in Figure 1. Source 1 becomes

> この法律は、施行の日から施行する。
> (*This Act shall come into force as from the date of enforcement*).

After transforming the source sentences, we applied hierarchical clustering. We used the group average method and the morpheme-based edit distance. The distance between two sentences is defined as the minimum number of operations needed to transform one sentence into the other, where an operation is the one of the insertion, deletion, or substitution of a single morpheme. However this distance is sensitive to the sentence length, so we normalized it into interval $[0, 1]$ by dividing by the sentence length.

We used the resulting clusters as PRTs except those containing only one sentence since such clusters are unreliable for evaluation.

Furthermore, fixed sentences are used in many statutes. For example, the sentence

[2] A *bunsetsu* is a linguistic unit in a Japanese sentence and roughly corresponds to a basic phrase in English.

この法律は、公布の日から施行する。
(*This Act shall come into force as from the day of promulgation.*)

appears in many statutes. From the viewpoint of consistent translation, the same source sentences should be translated into the same translation. If the same sentence is included more than once in the corpus, we use the translations of the sentence as reference translations instead of the cluster.

3.2. Modifying BLEU Metric

Since the BLEU metric considers both adequacy and fluency of the translation, we might easily consider it an evaluation metric for our purpose. However, there are some problems using PRTs. Thus we modified it as described in the following subsections.

3.2.1. Problems with BLEU Metric

The preceding section showed how to get pseudo reference translations. Although the reference translations used in the BLEU metric are the translations of a given source sentence, *pseudo* reference translations are not. This causes some evaluation problems. For example, consider the following two candidate translations:

Source 次に掲げる者は、委員となることができない。

Candidate 1. The following persons may not act as committee members:

Candidate 2. The following persons may not act as members:

For comparison, we prepared the following two PRTs:

Pseudo Reference 1. The following persons may not act as directors:
(次に掲げる者は、取締役となることができない。)

Pseudo Reference 2. The following persons may not act as accounting auditors:
(次に掲げる者は、会計監査人となることができない。)

Both Candidates 1 and 2 obviously resemble each other, and the only difference is the equivalent of "委員": "*committee members*" and "*members*." Both translations have identical adequacy, even though their BLEU scores are different. Particularly, p_1 is calculated by dividing the number of unigrams occurring in both a candidate and any its references by the number of unigrams occurring in the candidate. Therefore, p_1 of Candidate 1 is 7/9 since it contains nine unigrams and seven occur in the references. In the same way, p_1 of Candidate 2 is 7/8. The length of the equivalents of "委員" affects the scores, but that is not desirable since they do not occur in any references. In other words, *n*-grams occurring in the candidate but not in the PRTs may reduce the BLEU scores too much.

3.2.2. Introducing Weight

Since pseudo reference translations may not be translations of the source sentence, some *n*-grams occurring in the candidate translation may not occur in the PRTs, reducing the BLEU score, as shown in the above example.
This suggests that we should consider only the *n*-grams that occur in the PRTs. In addition, we assume that if an *n*-gram occurs in many PRTs, it may occur in the candidate translation. Therefore we introduce a weight $w(n\text{-}gram)$ that indicates the ratio of sentences containing *n-gram* and propose the following weighted BLEU metric (**BLEU-W**):

$$w(n\text{-}gram) = \frac{\#\text{ of sentences with } n\text{-}gram \text{ in PRTs}}{\#\text{ of sentences in PRTs}}, \tag{4}$$

$$p_n = \frac{\sum_{n\text{-}gram \in S} Count_{clip}(n\text{-}gram) * w(n\text{-}gram)}{\sum_{n\text{-}gram \in S} Count(n\text{-}gram) * w(n\text{-}gram)}, \tag{5}$$

$$\text{BLEU-W} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{6}$$

where BP is same as the equation (2).
The original BLEU metric evaluates the whole of translation documents, while the BLEU-W metric evaluates each sentence. By this modification, $w(n\text{-}gram)$ would be 0 if *n-gram* does not occur in the PRTs, since it does not contribute to the score.

3.2.3. Problems with BLEU-W Metric

Although the above BLEU-W metric succeeded in removing the negative effects of *n*-grams that occur only in the candidates, it causes another problem. Consider the following candidates to the source in Section 3.2.1.

Candidate 2. The following persons may not act as members:

Candidate 3. The following persons can not act as members:

The difference is "*may*" and "*can*." Since "*may*" is used in both Pseudo References 1 and 2 but "*can*" is not used in them, Candidate 2 is more appropriate than Candidate 3. Despite this, the BLEU-W metric does not consider "*can*" because it does not occur in the

PRTs, i.e., $w(\text{"}can\text{"}) = 0$. Thus p_1 of Candidate 2 is 7/7 and 6/6 for Candidate 3; they cannot be distinguished.

3.2.4. Considering Recall

To solve this problem, we should use more *n*-grams for the score calculation. In the BLEU-W metric, we assumed that an *n*-gram occurring in many PRTs *may* occur in the candidate translation and we did not consider the *n*-grams that do not occur in the candidate.

However, from the viewpoint of consistency, an *n*-gram occurring in many PRTs *should* occur in the candidate translation since the PRTs are the translations of sentences similar to the source sentence, and such *n*-grams may be fixed expressions.

Therefore we define $TopRef(n,\alpha)$ as the set of *n*-grams occurring more than the ratio $\alpha(0 \leq \alpha \leq 1)$ in the PRTs and use $TopRef(n,\alpha)$ for the score calculation as follows:

$$TopRef(n,\alpha) = \{n\text{-}gram \in Ref | w(n\text{-}gram) > \alpha\}, \tag{7}$$

where Ref is the set of *n*-grams in the PRTs. Using $TopRef(n,\alpha)$, we modify p_n as follows:

$$p_n = \frac{\sum\limits_{n\text{-}gram \in S \cup TopRef(n,\alpha)} Count_{clip}(n\text{-}gram) * w(n\text{-}gram)}{\sum\limits_{n\text{-}gram \in S \cup TopRef(n,\alpha)} \max(Count(n\text{-}gram),1) * w(n\text{-}gram)}, \tag{8}$$

$$\text{CIEL} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{9}$$

We call this metric **CIEL** (ConsIstency Evaluation for Legal documents). This metric can be considered a modification of BLEU with a recall-oriented strategy that came from the ROUGE metric (Lin, 2004).

3.3. Compliance Rate

Since pseudo reference translations may not be translations of source sentences, the BLEU-W and the CIEL metrics can evaluate only expressions occurring in the PRTs. For example, both metrics cannot evaluate the equivalents of “委員” in the Candidates 1 and 2 in Section 3.2.1. since the equivalents do not occur in the PRTs. Therefore both BLEU-W and CIEL are insufficient metrics for adequacy.

However, we can evaluate translation adequacy, i.e., whether adequate equivalents were used, using the Standard Bilingual Dictionary (SBD). The adequacy can be evaluated by considering standard equivalents in candidate translations. Thus we define compliance rate (CR) to evaluate adequacy:

$$\text{CR} = \frac{\sum\limits_{Source\ sentences} \text{\# of occurrences of SBD equivalents}}{\sum\limits_{Source\ sentences} \text{\# of occurrences of SBD entries}}. \tag{10}$$

4. Evaluation Experiment

We evaluated proposed metrics by comparing with other metrics through an experiment.

4.1. Experimental Targets

We evaluated the BLEU-W and the CIEL metrics by calculating the scores for the translations of the Labor Standards Act, which contains 242 sentences. We prepared three translations: by the Japanese government, a publishing company, and a Google translation tool. We calculated BLEU, BLEU-W, and CIEL scores for each translation as well as CR.

The government translation was done by legal specialists using the SBD[1] (Study Council, 2006). The company translation was also done by legal specialists, but not using the SBD. Since the government translation is based on the SBD, it is expected more consistent than the company one. The Google translation was the result of the machine translation system created by Google[3], so that it can be considered worst. Therefore the proposed metric is expected to highly score them in this order.

For the compilation of PRTs, we used the parallel corpus of Japanese statutes translated by the Japanese government[1], which include 17,793 Japanese sentence types of 84 statutes, excluding the Labor Standards Act and their translations as 20,154 English sentence types; one Japanese sentence sometimes has several English translated sentences.

In the CIEL metric, we set parameter α of $TopRef$ to 0.5.

4.2. Experimental Results

Before calculating the scores for the candidate translations, we divided the 17,793 Japanese sentences into 1,910 clusters and selected the closest cluster to each source sentence as mentioned in Section 3.1. But we could not calculate the score for 22 of the 242 sentences in the Labor Standards Act for the following reasons. In eight sentences, the closest cluster could not be selected since the last morphemes did not occur in any cluster. Other eight sentences had no closest cluster, since their distances to any clusters were equal

[3] http://www.google.com/translate_t

Distance to	BLEU			BLEU-W			CIEL		
the cluster	gov.	company	Google	gov.	company	Google	gov.	company	Google
$0.0 < d \leq 0.1$	0.352	0.322	0.126	0.852	0.869	0.559	0.551	0.451	0.157
$0.1 < d \leq 0.2$	0.307	0.320	0.109	0.821	0.784	0.425	0.435	0.423	0.104
$0.2 < d \leq 0.3$	0.298	0.299	0.125	0.738	0.675	0.427	0.351	0.351	0.131
$0.3 < d \leq 0.4$	0.153	0.143	0.076	0.588	0.579	0.273	0.262	0.231	0.079
$0.4 < d \leq 0.5$	0.138	0.116	0.069	0.518	0.483	0.233	0.185	0.140	0.081
$0.5 < d \leq 0.6$	0.089	0.087	0.063	0.344	0.285	0.215	0.117	0.109	0.059
$0.6 < d \leq 0.7$	0.074	0.069	0.059	0.331	0.286	0.190	0.110	0.101	0.066
$0.7 < d \leq 0.8$	0.076	0.072	0.061	0.300	0.282	0.268	0.112	0.094	0.079
$0.8 < d \leq 0.9$	0.054	0.052	0.071	0.128	0.128	0.188	0.048	0.046	0.043
$0.9 < d \leq 1.0$	0.059	0.058	0.053	0.093	0.093	0.093	0.053	0.055	0.051
Average	0.133	0.127	0.073	0.456	0.418	0.259	0.190	0.169	0.077

Table 1: Scores of BLEU, BLEU-W, and CIEL.

Distance to	BLEU			BLEU-W			CIEL		
the cluster	gov.	company	Google	gov.	company	Google	gov.	company	Google
$0.0 < d \leq 0.1$	2.80	2.56	1.00	1.52	1.55	1.00	3.51	2.87	1.00
$0.1 < d \leq 0.2$	2.82	2.94	1.00	1.93	1.85	1.00	4.19	4.07	1.00
$0.2 < d \leq 0.3$	2.39	2.40	1.00	1.73	1.58	1.00	2.67	2.67	1.00
$0.3 < d \leq 0.4$	2.02	1.89	1.00	2.16	2.12	1.00	3.32	2.93	1.00
$0.4 < d \leq 0.5$	1.99	1.68	1.00	2.22	2.07	1.00	2.28	1.73	1.00
$0.5 < d \leq 0.6$	1.40	1.38	1.00	1.60	1.33	1.00	1.98	1.85	1.00
$0.6 < d \leq 0.7$	1.25	1.17	1.00	1.75	1.51	1.00	1.65	1.52	1.00
$0.7 < d \leq 0.8$	1.24	1.17	1.00	1.12	1.05	1.00	1.42	1.19	1.00
$0.8 < d \leq 0.9$	0.76	0.74	1.00	0.68	0.68	1.00	1.10	1.06	1.00
$0.9 < d \leq 1.0$	1.10	1.10	1.00	1.00	1.00	1.00	1.05	1.08	1.00
Average	1.82	1.73	1.00	1.75	1.61	1.00	2.47	2.20	1.00

Table 2: Proportional Scores of BLEU, BLEU-W, and CIEL.

to 1, that is, there were no clusters similar to them. In the remaining six sentences, since their closest clusters contained only one sentence, they were unreliable.

Thus we calculated the BLEU, BLEU-W, and CIEL scores for the translations of the remaining 220 sentences, and the result is shown in Table 1.

As seen in the average values in Table 1, all three metrics scored in the expected order. However, the BLEU metric showed significant differences between the Google translation and the other two translations, but not between the government translation and the company one. On the contrary, the BLEU-W and the CIEL metrics showed significant differences between the government translation and the company one as well as between the Google one and the others.

In addition, the BLEU-W metric calculated identical scores for 53 of the 220 sentences; for these sentences it could not determine which translation was better. However, the CIEL metric only calculated fifteen sentences with identical scores, and seven sentences had identical translations between the government and the company.

When the distance between the source sentence and its closest cluster is small, the CIEL metric has a large difference between the government and the company. Table 2 shows the scores divided by each of the Google ones so that it makes clear the difference. From this, we conclude that the CIEL metric is reliable when the distance is small. However, the CIEL metric is unreliable when the distance is large, and we further discuss it in the following subsection.

On the other hand, the CR score, shown in Table 3, has also a desirable order.

4.3. Discussion

In the average score of the CIEL metric showed in Table 1, the government translation outperformed the company one. However, examination of each sentence

Translator	# of entries whose equivalents occur in the translation	CR
government	2,042	0.779
company	1,765	0.674
Google	1,533	0.585

(# of entries occurring in Labor Standards Act: 2,620)

Table 3: Compliance Rate.

reveals that some have undesirable results, meaning the company translation has a higher score than the government one. This is caused by the following two reasons.

First, when the distance between the source sentence and its closest cluster is small, the company translation *really* outperforms the government one. The CIEL metric gives a higher score to the better translation and this metric is desirable.

For example, as shown in Figure 2, the closest cluster of the source sentence

> "前項の委員会は、次の各号に適合するものでなければならない。"

has sentences that include the following fixed expression:

> "... の... は、次の各号に適合するものでなければならない。".

The government translation uses an equivalent "*must*" for "なければならない (*nakereba-naranai*)." However, the PRTs suggest that it should be translated not as "*must*" but "*shall*." Since the company translation uses "*shall*" in the PRTs, its CIEL outscores the government. In the same way, while "次の各号に (*tsugino-kakugouni*)" should be translated into "*with each of the following items*," the government translation uses "*to the following items*," reducing its CIEL score.

Despite this, the BLEU-W metric scored $p_n = 1$ $(n = 1,2,3,4)$ for both the government and the company translations; in this example, the BLEU-W metric cannot determine which is better.

Second, when the distance between the source sentence and its closest cluster is not small, the CIEL metric is unreliable since the source sentence does not resemble the sentences in the cluster. In such a case, the *n*-gram that should be used in the candidate translation does not occur in the PRTs and the *n*-gram contained in $TopRef(n,\alpha)$ does not occur in the candidate translation, either. As a result, the CIEL scores become small and the difference between the government and the company translation scores also becomes small, as shown in Table 1. So the company translation score sometimes becomes bigger than the government one. Therefore the CIEL metric cannot determine which is better if a source sentence has a large distance to its closest cluster. To solve this problem, we must collect a more reliable parallel corpus and make a cluster closer to a source sentence.

4.4. Comparison with Other Evaluation Metrics

To compare the CIEL metric with other evaluation metrics, we calculated 95% confidence intervals based on different samplings of the test data. This comparison is proposed for two machine translation systems (Zhang and Vogel, 2004).

First, create test suites $T_0, T_1, \ldots, T_B$, where T_1 to T_B are artificial test suites created by resampling T_0. Then, system X scored x_0 on T_0 and system Y scored y_0. The discrepancy between systems X and Y is $\delta_0 = x_0 - y_0$. Repeat this process on every B test suite and we have B discrepancy scores: $\delta_1, \delta_2, \ldots, \delta_B$. From these B discrepancy scores, find the middle 95% of the scores (i.e. the 2.5th percentile: $score_{low}$ and the 97.5th percentile $score_{up}$). $[score_{low}, score_{up}]$ is the 95% confidence interval for the discrepancy between machine translation systems X and Y. If the confidence interval does not overlap with zero, the difference between systems X and Y is statistically significant.

In our comparison, neither are machine translations, but X is the government translation of the Labor Standards Act, and Y is the company one. If an evaluation metric can claim a significant difference between the two translations, the metric is desirable.

We compared CIEL with Word Error Rate (WER), Position independent Word Error Rate (PER) (Leusch et al., 2003), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002), ROUGE (Lin, 2004) and BLEU.

The Labor Standards Act contains 242 sentences. However, from the above experiment, when the distance between a source sentence and its closest cluster exceeded 0.5, such a cluster is not reliable as PRTs. So we only used 94 sentences that have a closer cluster than a distance of 0.5. We also set $B = 2000$ as the number of times to repeat the process. Table 4 shows the results.

Since the confidence interval does not overlap with zero at PER, METEOR, and CIEL, only these three metrics can claim that the government and the company translations are significantly different. We also count the number of undesirable cases, which are the scores when the government translation is less than the company one, as shown in the most right column of

Source Sentence: 前項の委員会は、次の各号に適合するものでなければならない。

Government Translation: The committee set forth in the preceding paragraph must conform to the following items:

Company Translation: The committee mentioned in the preceding paragraph shall conform to each of the following items:

Pseudo Reference Translation 1: The statement of the detailed explanation of the invention as provided in item 3 of the preceding Paragraph shall comply with each of the following items:
(前項第三号の発明の詳細な説明の記載は、次の各号に適合するものでなければならない。)

Pseudo Reference Translation 2: The statement of the scope of claims as provided in paragraph 2 shall comply with each of the following items:
(第二項の実用新案登録請求の範囲の記載は、次の各号に適合するものでなければならない。)

Pseudo Reference Translation 3: The statement of the scope of claims as provided in paragraph 2 shall comply with each of the following items:
(第二項の特許請求の範囲の記載は、次の各号に適合するものでなければならない。)

Figure 2: An example where company translation outperforms government one.

Metric	$Score_{low}$	$Score_{up}$	# of undesirable cases
1 – WER†	-0.015	0.017	806
1 – PER†	0.004	0.022	4
METEOR	0.003	0.036	30
NIST	-0.000	0.021	68
ROUGE-1	-0.002	0.020	103
ROUGE-2	-0.135	0.208	707
BLEU	-0.005	0.026	162
BLEU-W	-0.016	0.067	230
CIEL	**0.012**	**0.060**	**2**

† We used the values subtracted from 1 for WEP and PER since they score lower if the translation is better, while other metrics score higher.

Table 4: Confidence intervals for discrepancy between two translations.

Table 4. The CIEL metric has the fewest cases, so it is the most desirable metric. Notice that the PER metric has few undesirable cases. However it only considers words and not word order; it only evaluates adequacy. Therefore the CIEL metric is the most desirable because it evaluates both adequacy and fluency.

5. Conclusion

In this paper, we proposed two consistency evaluation metrics for legal translations; the CIEL metric with pseudo reference translations and compliance rate CR with the Standard Bilingual Dictionary. The CIEL metric is based on *n*-gram alignment scoring and clustering algorithms, both of which are suitable for Japanese legal documents that contain recurrent phrasal structures.

We also confirmed that these metrics can evaluate several translations of one source sentence from the viewpoint of consistency.

Since the CIEL metric requires suitable pseudo reference translations, collecting consistent legal translations for them is future work. In addition, the CIEL metric is relative, that is, it determines which is better when several candidate translations are given. We also need an absolute metric that can determine whether one candidate translation is consistent. Furthermore, we want to examine how the CIEL metric correlates to the intuitive evaluation by human experts.

We intend to use the proposed metrics in the English Translation Project of Japanese Statutes[1] to determine whether the first translation of a statute is appropriate for the project database that aims for consistency and reliability of the translation.

6. Acknowledgements

The authors would like to thank Professor KASHIWAGI Noboru, Chuo Law School, Chair of Experts Council for Translation of Japanese Laws and Regulations into Foreign Languages, Ms. SHIMADA Chiko, Deputy Director, Cabinet Secretariat, and Mr. OHYA Tai, Prosecutor, Ministry of Justice, for their cooperation. Thanks are also owed to Professor MATSUURA Yoshiharu, Associate Professors KAKUTA Tokuyasu and Frank BENNETT, and Mr. SANO Tomoya, Graduate School of Law, Nagoya University, for their discussions and suggestions. This research project is partly supported by a Grant-in-Aid for Scientific Research (B) from the Japan Society for the Promotion of Science.

7. References

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgements. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Study Council for Promoting Translation of Japanese Laws and Regulations into Foreign Languages. 2006. Final report. http://www.cas.go.jp/jp/seisaku/hourei/report.pdf.

A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

G. Leusch, N. Ueffing, and H. Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings MT Summit IX*, pages 240–247.

C. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out, Workshop at the ACL'04*, pages 74–81.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

K. Toyama, Y. Ogawa, K. Imai, and Y. Matsuura. 2006. Application of word alignment for supporting translation of Japanese statutes into English. In T. M. van Engers, editor, *Legal Knowledge and Information Systems JURIX 2006: The Nineteenth Annual Conference*, pages 141–150.

Y. Zhang and S. Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 85–94.

Automatic Summarization and Information Extraction from Canadian Immigration Decisions

Emmanuel Chieze[*], Atefeh Farzindar[#], Guy Lapalme[*]

[*]RALI-DIRO
C.P. 6128, Succ Centre-Ville
Université de Montréal
Montréal, Québec, Canada H3C 3J7

[#]NLP Technologies
3333, chemin Queen Mary, suite 543,
Montréal, Québec, Canada H3V 1A2
E-mail: chiezeem@iro.umontreal.ca, farzindar@nlptechnologies.ca, lapalme@iro.umontreal.ca

Abstract

This paper presents our experience in the use of a mix of linguistics aware transductor and XML technologies for bilingual information extraction from judgments in both French and English within a legal information and summarizing system. We present the context of the work, the main challenges and how they were tackled by clearly separating language and domain dependent terms and vocabularies. The use of Excel sheets for keeping dictionary information enables an easy to use customization approach for non linguists or non computer scientists.

1. Context of the work

One field in which information is produced in large quantities and needs to be adequately classified and be reliably accessible is the legal field. Indeed, legal experts perform relatively difficult legal clerical work that requires accuracy and speed. These legal experts often summarize legal documents, such as court judgments, and look for information relevant to specific cases in these summaries. These tasks involve understanding, interpreting, explaining and researching a wide variety of legal documents. The summary of a judgment is a compressed but accurate statement of the judgment's contents. Summaries help organize a large volume of documents so that finding relevant judgments for a specific case is easy and efficient.

That is why judgments are frequently manually summarized by legal experts. However, human time and expertise required to provide manual summaries for legal research make human-generated summaries relatively expensive. Also, there is always the risk that a legal expert misinterprets a judgment and misclassifies it or produces an erroneous summary.

Because of the high accuracy required in the classification and summarization of legal judgments, commonly available automatic classification and summarization methods are typically not suitable for this task.

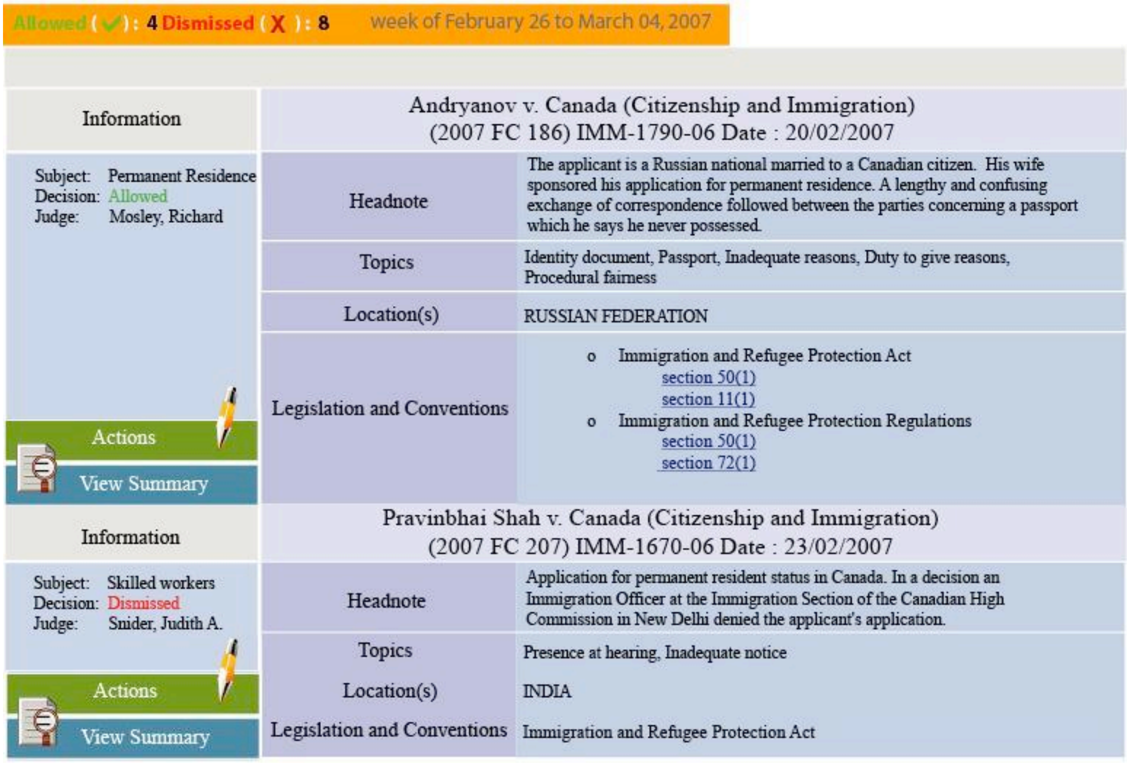
NLP Technologies is an enterprise that develops solutions specifically for users of legal search tools. The company's services are available through the company's website[1] and include access to four main tools:

- *DecisionExpress* is the tool that processes judicial decisions automatically and makes the information used by jurists daily more accessible by presenting the summaries of the legal record of the proceedings of federal courts in Canada as a table-style summary as shown in Figure 1. This service provides some form of continuing education for legal practitioners and saves hours of reading by extracting the essential information and showing it in a uniform format for many cases of the same type.
- *SearchExpress,* integrated within *DecisionExpress* is a search engine that allows users to search the NLP Technologies' database.
- *BiblioExpress* is a virtual law library providing access to legislations, regulations and international instruments.
- *StatisticExpress* is a specialized fact-finder providing fast and easy access to pertinent data and government statistics.

Since August 2006, the Federal Court of Canada has been using the services of NLP Technologies. The summaries are available within 2 days of the publication date.

FLEXICON (Smith & Deedman, 1987), SALOMON (Moens et al.,1999) and SUM projects (Grover et al.,2003) attest the importance of the exploration of legal knowledge for sentence categorisation and summarization (Moens 2007). NLP Technology's extraction of the most important units is based on the identification of the thematic structure in the document and the determination of argumentative themes of the textual units in the judgment (Farzindar & Lapalme, 2004; Farzindar,2005).

[1] http://www.NLPtechnologies.ca



Allowed (✓): 4 Dismissed (X): 8 week of February 26 to March 04, 2007

Information	Andryanov v. Canada (Citizenship and Immigration) (2007 FC 186) IMM-1790-06 Date : 20/02/2007	
Subject: Permanent Residence Decision: Allowed Judge: Mosley, Richard	Headnote	The applicant is a Russian national married to a Canadian citizen. His wife sponsored his application for permanent residence. A lengthy and confusing exchange of correspondence followed between the parties concerning a passport which he says he never possessed.
	Topics	Identity document, Passport, Inadequate reasons, Duty to give reasons, Procedural fairness
	Location(s)	RUSSIAN FEDERATION
	Legislation and Conventions	o Immigration and Refugee Protection Act section 50(1) section 11(1) o Immigration and Refugee Protection Regulations section 50(1) section 72(1)
Information	Pravinbhai Shah v. Canada (Citizenship and Immigration) (2007 FC 207) IMM-1670-06 Date : 23/02/2007	
Subject: Skilled workers Decision: Dismissed Judge: Snider, Judith A.	Headnote	Application for permanent resident status in Canada. In a decision an Immigration Officer at the Immigration Section of the Canadian High Commission in New Delhi denied the applicant's application.
	Topics	Presence at hearing, Inadequate notice
	Location(s)	INDIA
	Legislation and Conventions	Immigration and Refugee Protection Act

Figure 1: *Factsheet* from *DecisionExpress* showing two cases from a week in which 4 immigration cases have been allowed and 8 dismissed. The left part give the subject, the decision and the name of the judge while the right part gives a very short summary, the topics dealt in this case, the country in which the applicant resided and the pertinent legislation that was cited in the case. By merely clicking on the appropriate button, it is possible to get a longer summary (shown in Figure 2) or even the text of the original judgment.

2. The *Immigration and Refugee Law*

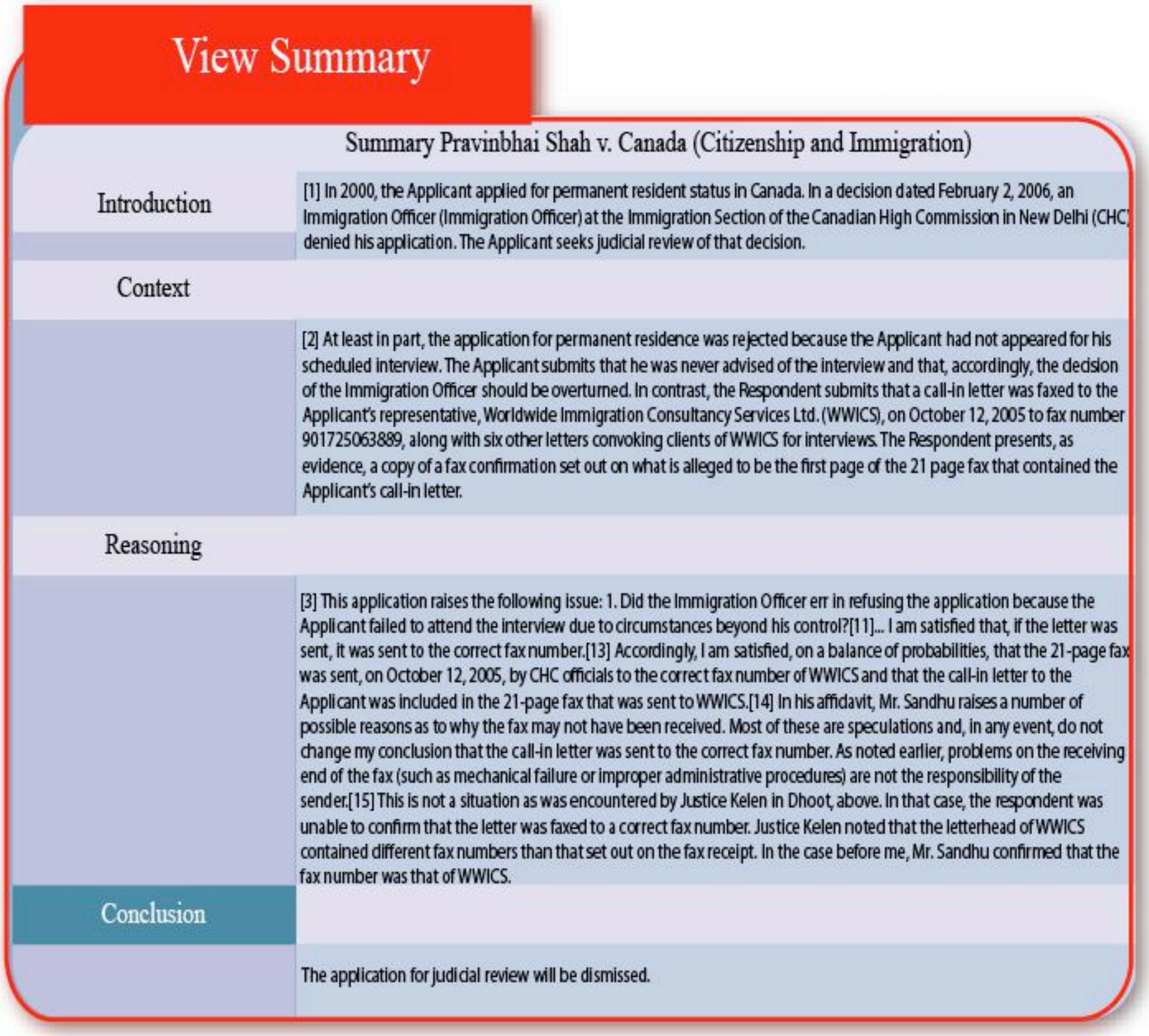
We describe in more detail the process of dealing with decisions in the field of Immigration and Refugee Law. All Canadian immigration decisions are retrieved from the Federal courts web site when they become public, and are then processed in order to produce two valuable pieces of information (See Figure 1),: a *Factsheet*, which is a fixed set of structured information automatically extracted from the decisions (name of the judge, name of the case, docket number and neutral citation number, place of hearing …), and an automatic summary of the decisions, a sequence of relevant sentences taken directly from the original decision and presented in a table (Figure 2). The factsheet clearly identifies such salient information as the subject matter, key words, presiding judge, result, legislation cited, etc., as well as an automatic summary composed of extracts from the decision and presented in a thematic table.

As the Court decisions in this domain are well structured, it is possible to identify three main parts and develop a specialized information extraction process for each:

1. **Prologue**: a list of semi-structured information such as the docket number, the place and date of hearing, the judge's, plaintiffs' and defendants' names. Each piece of information is usually introduced by a specific label but the concept extraction and the determination of the matter of the decision require a more detailed analysis.
2. **Decision**: a full-length text, structured in sections usually identified by titles or by specific sentences starting those sections. A typical decision is divided into six themes usually appearing in the following order: introduction, context, issues raised by the plaintiffs, reasoning, conclusion and the order. Some sections may be missing in some decisions, while additional sections may appear in other ones. The order in which sections appear may also vary.
3. **Epilogue**: another list of semi-structured information such as the lawyers' and solicitors' names.

The information from the prologue and epilogue are kept in a database and an automatic table style summary is produced for the decision. The result is then reviewed by a lawyer from NLP Technologies who can make some manual adjustments. The overall result is revised by an Editorial Board before the information becomes available to the company's subscribers on the Web. This mix of automatic processing and manual revision has been in operation for 2 years and has given very good results on Immigration decisions written in English.

We now describe a new version of that process, being gradually put in production, to extend the system to decisions in the same field written in French and to decisions in other fields covered by the Canadian Federal Law such as tax law and intellectual property law. Two core ideas have presided to this re-engineering: the use of linguistics aware technology and parameterization.



Summary Pravinbhai Shah v. Canada (Citizenship and Immigration)

Introduction

[1] In 2000, the Applicant applied for permanent resident status in Canada. In a decision dated February 2, 2006, an Immigration Officer (Immigration Officer) at the Immigration Section of the Canadian High Commission in New Delhi (CHC) denied his application. The Applicant seeks judicial review of that decision.

Context

[2] At least in part, the application for permanent residence was rejected because the Applicant had not appeared for his scheduled interview. The Applicant submits that he was never advised of the interview and that, accordingly, the decision of the Immigration Officer should be overturned. In contrast, the Respondent submits that a call-in letter was faxed to the Applicant's representative, Worldwide Immigration Consultancy Services Ltd. (WWICS), on October 12, 2005 to fax number 901725063889, along with six other letters convoking clients of WWICS for interviews. The Respondent presents, as evidence, a copy of a fax confirmation set out on what is alleged to be the first page of the 21 page fax that contained the Applicant's call-in letter.

Reasoning

[3] This application raises the following issue: 1. Did the Immigration Officer err in refusing the application because the Applicant failed to attend the interview due to circumstances beyond his control?[11]... I am satisfied that, if the letter was sent, it was sent to the correct fax number.[13] Accordingly, I am satisfied, on a balance of probabilities, that the 21-page fax was sent, on October 12, 2005, by CHC officials to the correct fax number of WWICS and that the call-in letter to the Applicant was included in the 21-page fax that was sent to WWICS.[14] In his affidavit, Mr. Sandhu raises a number of possible reasons as to why the fax may not have been received. Most of these are speculations and, in any event, do not change my conclusion that the call-in letter was sent to the correct fax number. As noted earlier, problems on the receiving end of the fax (such as mechanical failure or improper administrative procedures) are not the responsibility of the sender.[15] This is not a situation as was encountered by Justice Kelen in Dhoot, above. In that case, the respondent was unable to confirm that the letter was faxed to a correct fax number. Justice Kelen noted that the letterhead of WWICS contained different fax numbers than that set out on the fax receipt. In the case before me, Mr. Sandhu confirmed that the fax number was that of WWICS.

Conclusion

The application for judicial review will be dismissed.

Figure 2: Automatically generated, and manually revised, summary returned after clicking on the *View Summary* button at the bottom left of Figure 1. All sentences of the summary being taken verbatim from the original decision, they can thus be used as argumentation. The sentences are classified into meaningful sections: Introduction, Context, Reasoning and Conclusion. Note that sentences are not necessarily in the same order in the judgment and in the summary.

3. Overview of the linguistics aware Information Extraction Process

Canadian immigration decisions are available on the Web[2] as HTML documents and can be in English or French depending on the language used at the hearing. A decision may naturally be relevant for Canadian lawyers no matter in which language it is written. Since HTML tags define the presentation of those decisions, rather than their structure, and since the presentation as well as its HTML definition is liable to evolve over time (and it has…), we cannot rely on only these tags to identify the structure of the decisions. We will thus have to analyze the text of the decision itself to discover what parts of the text are part of each section that will appear in the summary.

Figure 3 shows a simplified view of the overall transformation pipeline combining different technologies to go from an original judgment as an HTML file taken from the web site of the Canadian Federal Court to an XML file that is saved within a data base from which the final summary, also in HTML, is generated. This XML file can also be changed during the manual revision process by NLP Technologies lawyers that access it through a specialized revision interface.

This transformation process involves both local (within a sentence) processing, more global processing taking into account parts of the documents that can be farther apart and statistical processing for computing the salient sentences that will compose the final summary.

We have decided to use technologies that are appropriate for each step of the transformation. Transductors allow a great flexibility in sentence processing, XSLT stylesheets are an efficient mean for selecting and transforming longer spans of texts and a procedural language is used for computing the final statistics to select the final sentences to appear in the summary.

[2] http://decisions.fct-cf.gc.ca/fr/index.html

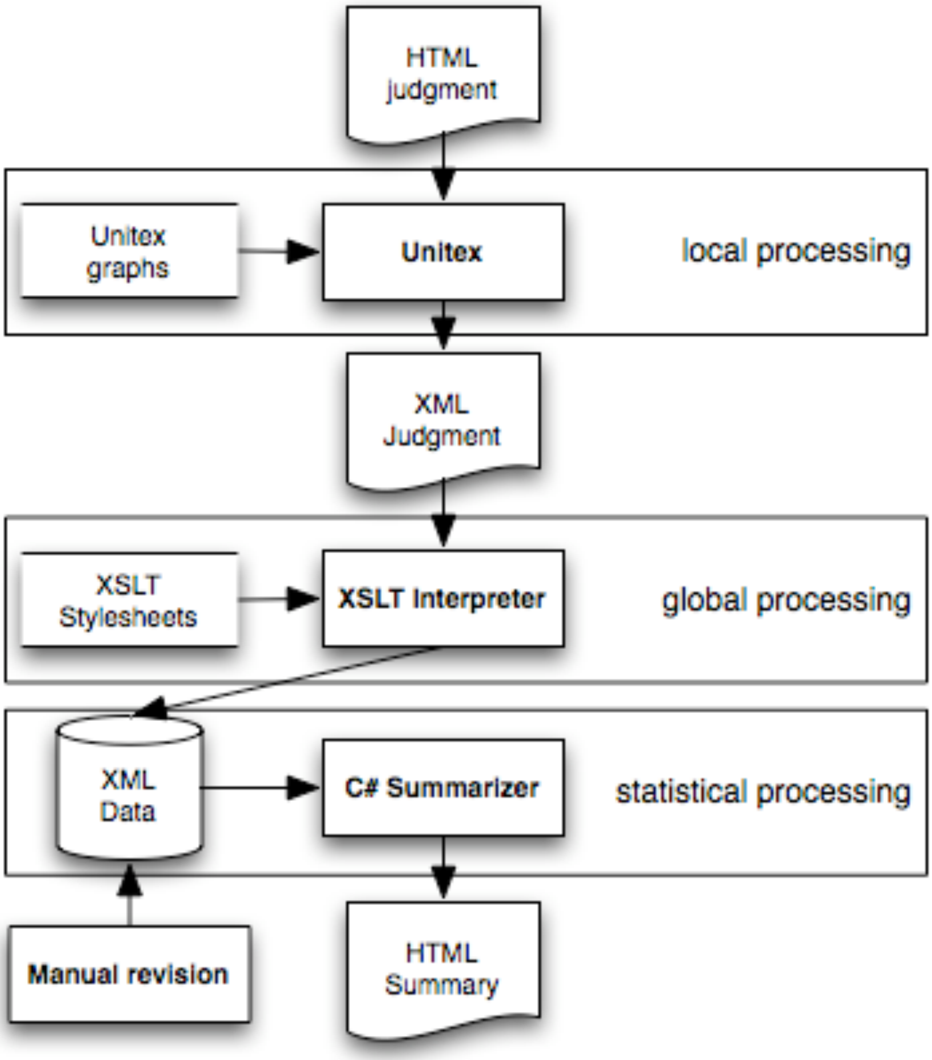


Figure 3: System architecture going from the original to the summary. Unitex graphs are used for going from HTML to XML and for linguistic processing within a sentence or for short spans of text. XML Transformation Stylesheets enable to take into account long distance dependencies and the statistical computations for determining the most important sentences to appear in the summary are done by a C# program.

3.1 Local processing

A first step is thus to convert HTML documents into text files and then use linguistic cues to identify the decision structure as well as the relevant factual information. Fortunately, decisions follow a rather stereotypical pattern and use recurrent information identifiers or section headings. Such identifiers have several variants, but there are usually a fixed set of them.

We decided to use XML tags to identify text structure and relevant factual information, since there are several general-purpose XML-based processing tools, such as structure validation or document transformation tools. So our process will first eliminate most HTML tags and transform others into paragraph markers.

Relevant information will then be identified through linguistic cues, which are phrases identifiable through context-free grammars. As we are aiming for power and flexibility we decided to make use of the transductor technology, namely Unitex[3], a descendant of INTEX (Silberztein 1973), to identify, mark and transform spans of texts by means of regular expressions which provides the following advantages:

1. Regular expressions are represented with graphs (see Figure 4 for an example) instead of complex sequences of operators and their base unit is the word rather than the character. Language-dependant character equivalences are appropriately handled.

[3] http://www-igm.univ-mlv.fr/~unitex/

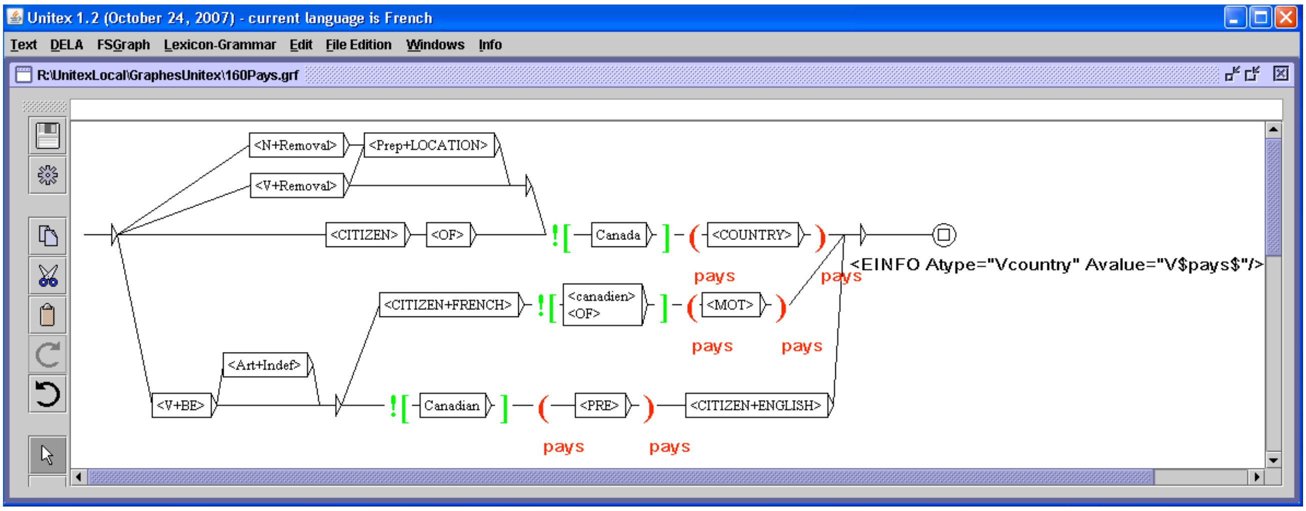


Figure 4: A graph defines a set of paths matching words encountered in the text going from the entry node (the triangle on the far left) to the exit node (the circle containing a square) on the right. A node can match either be a single word (see *Canadian* above), or one word contained in a list defined in the dictionary (see <*COUNTRY*> above). When a path going from the entry to exit has been found, information can be added (shown here in bold) to the original text. Here the occurrence detected is tagged with an XML tag named *EINFO* with attributes *ATYPE* having value *country* and *Avalue* having a value *pays* that was saved during matching this graph. This graph detects the country from which the applicant originates. The 4 paths out of the start state, from top to bottom, correspond respectively to: 1) a path that recognized phrases such as "his removal to Kenya", 2) a path that recognizes phrases such as "[is scheduled to be] removed to Kenya", 3) a path that recognizes phrases such as "[is a] citizen of Kenya", 4) a path that recognizes phrases such as "[is a] Kenyan citizen" or "[est] citoyen kenyan". Note that adjectives derived from country names, recognized by the last path, are not listed in the dictionary contrary to country names, which are listed.

2. It works with a user-defined dictionary in which words and phrases may be assigned various user-defined syntactic or semantic categories which may in turn be used in graphs. Flexional categories and morphological criteria can be almost freely combined with those syntactic and semantic categories, enabling the expression of complex search criteria without ever having to translate those criteria into character patterns.

3. Graphs may be used as subgraphs of other more complex graphs, enabling graph reusability.

4. Parameterized graphs (explained in the next section) add even more flexibility to our processing.

Unitex graphs have the power and efficiency of regular expressions, with the additional benefits of linguistic awareness and much improved user-friendliness. These grammars recognize word patterns most often limited to a single sentence. Unitex processing of the judgments involve the use of 33 compiled graphs for transforming the HTML form of a judgment to a labeled XML file. An example of such a graph that detects the applicant's country of origin is displayed as Figure 4.

3.2 Global processing

Although there is no theoretical bounds on the span of input that can be processed by a Unitex transductor, in practice we have experienced many problems when the input is too long. Unitex is cumbersome for expressing long-range dependencies but there are however a few contextual or structural rules to implement, such as:

- A sentence that contains a pattern associated with salient phrases of section X is a salient phrase of section X if and only if it appears in that section.

- All sentences of a paragraph following a sentence identified as a citation are also part of that citation.

We decided to express such structural rules with XSLT stylesheets applied to the resulting XML format of the documents.

Using XML provides the additional benefit of checking the conformity of the document structure to the XML schema associated with decisions. The XSLT processing uses 10 templates.

3.3 Statistical processing

The above processing has tagged the original text without modifying it but to identify the sentences to appear in the summary, some statistical computations are involved such as the computation of TF•IDF scores and other numerical values. This process is done with a C# program that parses the XML document produced by the previous two steps.

The HTML input files are about 30K characters long, corresponding to 16K words. On a stock desktop PC, the processing time for applying Unitex graphs, processing XSLT templates and computing statistics is about 40 seconds by judgment.

4. Parameterization of the Information Extraction Process

As shown in Figure 4, Unitex graphs can refer to words defined in a dictionary, a user-defined list of word forms associated with their root form as well as various syntactic and semantic categories and morphological features. It would be cumbersome to define all word forms by hand, especially in an inflected language like French in which semantic categories do not vary with the flexion, Unitex offers two types of dictionary definitions: the inflected dictionary, where it is possible to directly define word forms, and the non inflected dictionary, which will be inflected by Unitex using an inflexion graph provided by the user. Such graphs are language dependent but are application domain independent.

Unitex offers an additional mechanism called the parameterized graph, which combines a generic graph containing variables and a parameter file. The latter is a text file containing the values to be taken by the variables. More precisely, each line of the parameter file will generate a subgraph, and the whole family of subgraphs will be integrated as a single graph. Each subgraph thus represents an alternative and the main graph a disjunction of all those alternatives.

In order to maximize the parameterization of our system, we have made an extensive use of the dictionary as well as of parameterized graphs, so that many graph updates can simply be made through the update of those parameter files followed by a graph recompilation. We have used Microsoft Excel in order to gather the various parameter files in one single place, to make data definition user-friendlier, to validate it with an Excel macro, which includes cross-checks between those lists when applicable and to give the user a user-friendly way of consulting, sorting and filtering those parameter lists.

Some operators such as `X` *`in-same-sentence-as`* `Y` or `X` *`near`* `Y`, not available in Unitex have been developed with auxiliary graphs, and can be used in those lists to implement complex rules: there is a fixed list of them however, since we did not want to implement a general rule compiler.

In total, there are 10 worksheets in this Excel file: each of them parameterizes a specific aspect of the information extraction process. The dictionary itself contains 432 uninflected single words, 840 inflected single words or single words without any flexion and 812 phrases. Those figures combine both English and French entries. In a specialized information extraction setting like this one, we only have to deal with words that are used for segmenting the judgment or for identifying specific information like dates, names of parties. Most of the words encountered in the text are simply taken as is and will be given back verbatim if it happens that the sentence as a whole is chosen to appear in the summary.

5. Maintenance of the Information Extraction Process

The information extraction transductors were developed originally by the manual inspection of about 60 decisions in both English and French published in 2007. Only a few (about 5%) of current decision are not processed correctly and imply some manual adjustment either by correcting the formatting of the input or by adding new words in the dictionary.

We have also tested the transductors on 14 380 *historical* decisions published between 1997 and 2006. Only 15% of those decisions were incorrectly processed by the original information extraction process, i.e. the resulting XML document was not well-formed, usually because the beginning of a section was detected but not its end or vice-versa. This happens because these complementary elements are tagged independently.

Resolving the problems caused by 9 decisions helped resolve the problems encountered in 49 additional decisions (over 90 decisions tested). In other words, a single problem occurred on average on 6.5 decisions among the 90 decisions on which corrections were tested. Among those 9 problems, 3 implied adding entries in the dictionary, 5 implied modifying existing graphs in order to improve their flexibility. We decided not to take any step on the last one which was caused by a misspelling in the decision. It is yet unclear whether our parameterization effort has been sufficient, since only 3 problems out of 8 could be solved without modifying any graph. We are just at the beginning of the correction process however, and we hope that, as time goes on, a higher proportion of problems will be solved through dictionary update, as well as we can hope that one single correction will have a positive impact on more decisions. Moreover, we know that decisions have been presented in a considerably more homogeneous way since 2003, so that historical results are worse than those obtained on current decisions.

So we are confident that as the time goes on, there will be less and less manual work to do by NLP technology legal staff who will merely check that everything is all right for publication. It is too early to do a formal evaluation of the new process both in terms of the efficiency of extraction and in the reduction of manual corrections needed before the judgment is put on the NLP Technologies web site. But as the process is good enough to be gradually put in production, we are very pleased with the result.

6. Conclusion and Perspectives

DecisionExpress is the first service in the world based on an automatic summarization system developed specifically for legal documents. It is implemented in a real-life environment and currently produces summaries for large collections of judgments (between 25 and 50 each week) written in English in the immigration domain.

In this article, we have presented our recent work for extending the applicability of the system to French and to other domains such as financial field and intellectual property field. The main idea was to separate the linguistic cues used to achieve a precise information extraction in different domains. The output of the system is systematically reviewed by a lawyer but the goal is to have the system do as much work as possible.

To allow NLP Technologies client to work in the language they are most comfortable with, a project of automatic translation summary of judgments is under way. That would help users during the (up to nine) months it takes for the official translation to be published. As the summaries are obtained with extracts of the original judgment, the decision could be summarized both in English and in French, regardless of the original language of the judgment by taking the corresponding extracts from the automatic translation.

7. Acknowledgement

This work is partially funded by the PRECARN Alliance Program of Canada and The National Research Council Canada (NRC).

8. References

Farzindar, A. (2005). *Automatic summarization of legal texts*, Ph.D. Thesis, University of Montreal and University of Paris IV-Sorbonne.

Farzindar, A. and Lapalme, G. (2004). LetSUM, an automatic Legal Text Summarizing System. In Thomas F. Gordon (editors), *Legal Knowledge and Information Systems, Jurix 2004: the Sevententh Annual Conference*, IOS Press, Berlin, pp. 11-18.

Grover, C., Hachey, B. and Korycinski, C. (2003). Summarising legal texts: Sentential tense and argumentative roles. In Radev, D. and Teufel, S., editors, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, Edmonton, Alberta, Canada, pp. 33–40.

Moens, M.F. (2007) Summarizing court decisions, *Information Processing and Management*, vol 43, pp. 1748-1764.

Moens, M.F., Uyttendaele, C. and Dumortier, J. (1999). Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2), pp. 151–161.

Silberztein, M.D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson, 234 p.

Smith, J.C. and Deedman, C. (1987). The application of expert systems technology to case-based law. *ICAIL*, pp. 84–93.