eXtended WordFrameNet

Egoitz Laparra, German Rigau

IXA NLP Group, University of the Basque Country, Donostia, Basque Country
{egoitz.laparra, german.rigau}@ehu.es

Abstract

This paper presents a novel automatic approach to partially integrate FrameNet and WordNet. In that way we expect to extend FrameNet coverage, to enrich WordNet with frame semantic information and possibly to extend FrameNet to languages other than English. The method uses a knowledge-based Word Sense Disambiguation algorithm for matching the FrameNet lexical units to WordNet synsets. Specifically, we exploit a graph-based Word Sense Disambiguation algorithm that uses a large-scale knowledge-base derived from existing semantic resources. We have developed and tested additional versions of this algorithm showing substantial improvements over state-of-the-art results. Finally, we show some examples and figures of the resulting semantic resource.

1. Introduction

Building large and rich predicate models for broad–coverage semantic processing as FrameNet (Baker et al., 1998), VerbNet (Kipper, 2005) or PropBank (Palmer et al., 2005) takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, the coverage of currently available predicate-argument resources is still unsatisfactory. For example, (Burchardt et al., 2005) or (Shen and Lapata, 2007) indicate the limited coverage of FrameNet as one of the main problems of this resource. Currently, FrameNet1.3 covers around 10,000 lexical-units while for instance, WordNet3.0 contains 206,941 word senses. Furthermore, the same effort should be invested for each different language (Subirats and Petruck, 2003). Following the line of previous works, we empirically study a novel approach to partially integrate FrameNet (Baker et al., 1998) and WordNet (Fellbaum, 1998). The method relies on the use of a knowledge–based Word Sense Disambiguation (WSD) algorithm that uses a large-scale graph of concepts derived from WordNet (Fellbaum, 1998) and eXtented WordNet (Mihalcea and Moldovan, 2001). The WSD algorithm is applied to coherent groupings of words belonging to the same frame. In that way we expect to extend the coverage of FrameNet (by including closely related concepts from WordNet), to enrich WordNet with frame semantic information (by porting frame information to WordNet) and to extend FrameNet to languages other than English (by exploiting local wordnets aligned to the English WordNet).

WordNet [1] (Fellbaum, 1998) is by far the most widely-used knowledge base. It contains manually coded information about English nouns, verbs, adjectives and adverbs and is organized around the notion of a *synset*. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, <*premier*, *prime_minister*, *chancellor*> form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss, in this case: ”the person who is head of state (in several countries)” and by explicit semantic relations to other synsets, including hypernymy/hyponymy, meronymy/holonymy, antonymy, entailment, etc.

FrameNet [2] (Baker et al., 1998) is a very rich semantic resource that contains descriptions and corpus annotations of English words following the paradigm of Frame Semantics (Fillmore, 1976). In frame semantics, a Frame corresponds to a scenario that involves the interaction of a set of typical participants, playing a particular role in the scenario. FrameNet groups words or Lexical Units (LUs hereinafter) into coherent semantic classes or frames, and each frame is further characterized by a list of participants or Frame Elements (FEs hereinafter). Different senses for a word are represented in FrameNet by assigning different frames.

Currently, FrameNet represents more than 10,000 LUs and 825 frames. More than 6,100 of these LUs also provide linguistically annotated corpus examples. However, only 722 frames have associated a LU. From those, only 9,360 LUs[3] where recognized by WordNet (around 92%) corresponding to only 708 frames.

LUs of a frame can be nouns, verbs, adjectives and adverbs representing a coherent and closely related set

[1] `http://wordnet.princeton.edu/`

[2] `http://framenet.icsi.berkeley.edu/`

[3] Word-frame pairs

of meanings that can be viewed as a small semantic field. For example, the frame LEADERSHIP contains LUs referring to the leadership activity and their participants. It is evoked by LUs like *leader.n*, *premier.n*, *government.n*, *lead.v*, *govern.v*, etc. The frame also defines core semantic roles (or FEs) such as GOVERNED or LEADER that are semantic participants of the frame. Note that some FEs also correspond to LUs associated to frame (see example below).

[Hussein]$_{LEADER}$ *governed* [Jordan]$_{GOVERNED}$.

The paper is organized as follows. After this short introduction, in section 2. we present a brief summary of the method used to integrate WordNet and FrameNet and the evaluation made in previous works. Section 3. shows how we build a new multilingual resource that extends the coverage of the LexicalUnits of FrameNet, and finally, in section 4., we draw some final conclusions and outline future work.

2. Building WordFrameNet

WordFrameNet (Laparra et al., 2010) is a new resource that combines knowledge from FrameNet and WordNet. In order to connect both resources we used a knowledge-based Word Sense Disambiguation algorithm for assigning appropriate WordNet synsets to the FrameNet lexical units. Specifically, we exploit a graph-based Word Sense Disambiguation algorithm called SSI-Dijkstra+, (Laparra et al., 2010) that is an advanced version of the Structural Semantic Interconnections algorithm (SSI)(Navigli and Velardi, 2005). SSI is a very simple algorithm consisting on an initialization step and a set of iterative steps.

Given W, an ordered list of words to be disambiguated, the SSI algorithm performs as follows. During the initialization step, all monosemous words are included into the set I of already interpreted words, and the polysemous words are included in P (all of them pending to be disambiguated). At each step, the set I is used to disambiguate one word of P, selecting the word sense which is *closer* to the set I of already disambiguated words. Once a sense is disambiguated, the word sense is removed from P and included into I. The algorithm finishes when no more pending words remain in P.

As SSI-Dijkstra (Cuadros and Rigau, 2008), in order to measure the proximity of one synset (of the word to be disambiguated at each step) to a set of synsets (those word senses already interpreted in I), SSI-Dijkstra+ uses as a knowledge base a very large connected graph with 99,635 nodes (synsets) and 636,077 edges (the set of relations between synsets gathered from WordNet[4] (Fellbaum, 1998) and eXtended WordNet[5] (Mihalcea and Moldovan, 2001). For building this graph we used WordNet version 1.6 and the semantic relations appearing between synsets and disambiguated glosses of WordNet 1.7. To map the relations appearing in eXtended WordNet to WordNet version 1.6 we used the automatic WordNet Mappings[6] (Daudé et al., 2003). SSI-Dijkstra+ uses the Dijkstra algorithm to obtain the shortest path distance between a node and some nodes of the whole graph. The Dijkstra algorithm is a greedy algorithm that computes the shortest path distance between one node an the rest of nodes of a graph. BoostGraph[7] library can be used to compute very efficiently the shortest distance between any two given nodes on very large graphs. On that graph, SSI-Dijkstra computes several times the Dijkstra algorithm.

Initially, the list I of interpreted words should include the senses of the monosemous words in W, or a fixed set of word senses. Note that when disambiguating a Lexical Unit to a particular synset, the list I always includes since the beginning at least the sense of the LU and the rest of monosemous words of W. However, many frames only group polysemous LUs. In fact, a total of 190 frames (around 26%) only have polysemous LUs. Thus, SSI-Dijkstra provides no results when there are no monosemous terms in W. In this case, before applying SSI, the set of the LUs corresponding to a frame (the words included in W) have been ordered by polysemy degree. That is, the less polysemous words in W are processed first.

Obviously, if no monosemous words are found, we need to adapt the SSI algorithm. In order to make an initial guess, we devised four different options trying to initialice the set I with the most probable sense of the less ambiguous word of W. These four different versions of the algorithm are explained in depth in (Laparra and Rigau, 2009) and (Laparra et al., 2010).

We have evaluated the performance of the different versions of the SSI-Dijkstra algorithm using the same data set used by (Tonelli and Pianta, 2009). This data set consists of a total of 372 LUs corresponding to 372 different frames from FrameNet1.3 (one LU per frame). Each LUs have been manually annotated with the corresponding WordNet 1.6 synset. This Gold Standard includes 9 frames (5 verbs and 4 nouns) with

[4] `http://wordnet.princeton.edu`

[5] `http://xwn.hlt.utdallas.edu`

[6] `http://www.lsi.upc.es/~nlp/tools/mapping.html`

[7] `http://www.boost.org/doc/libs/1_35_0/libs/graph/doc/index.html`

only one LU (the one that has been sense annotated). Obviously, for these cases, our approach will produce no results since no context words can be used to help the disambiguation process[8].

As expected, the SSI-Dijkstra algorithms present different performances according to the different POS (Laparra and Rigau, 2009) and (Laparra et al., 2010). Also as expected, verbs seem to be more difficult than nouns and adjectives as reflected by both the results of the baseline and the SSI-Dijkstra algorithms.

As a result of the empirical study presented in (Laparra et al., 2010), we developed **SSI-Dijkstra+** a new version of SSI-Dijkstra combining the strategies of the two versions of the algorithm that perform better.

Table 1 presents detailed results per Part-of-Speech (POS) of the performance of the SSI-Dijsktra+ algorithm on the Gold Standard in terms of Precision (P), Recall (R) and F1 measure (harmonic mean of recall and precision). As baseline, we also include the performance measured on this data set of the most frequent sense according to the WordNet sense ranking (*wn-mfs*). Remember that this baseline is very competitive in WSD tasks, and it is extremely hard to beat upon even slightly (McCarthy et al., 2004). In order to show the improvement over the original SSI-Dijkstra algorithm we also include in this table the results obtained by in the same dataset. Notice that the original SSI-Dijkstra algorithm achieves a higher precision but a lower recall than SSI-Dijkstra+, because of the frames containing only polysemous words.

Table 2 presents detailed results of the performance of the SSI-Disjktra+ algorithm on the FrameNet–WordNet Verbal mapping (*VM*) produced by (Shi and Mihalcea, 2005) in terms of Precision (P), Recall (R) and F1 measure. Again, we include on the results obtained by the original SSI-Dijsktra algorithm in this dataset and also the most frequent sense according to the WordNet sense ranking (*wn-mfs*).

On this dataset, the overall results are much higher because this dataset provides several correct verbal senses per LU. Again, the knowledge-based WSD algorithms perform over the most frequent sense baseline.

In fact, we expect much better results performing the disambiguation process including in I, when available, the manually assigned FrameNet–WordNet Verbal mappings. Possibly, using this approach very high accuracies for nouns, adjectives and the remaining verbs could be obtained.

[8]In fact, FrameNet has 33 frames with only one LU, and 63 with only two.

	P	**R**	**F**
mfs-wn	0.67	0.67	0.67
SSI-Dijkstra	**0.79**	0.74	0.76
SSI-Dijkstra+	**0.79**	**0.79**	**0.79**

Table 2: Results using FN–WN Verbal mapping from (Shi and Mihalcea, 2005) as gold standard

3. Building eXtended WordFrameNet

After the the integration of WordNet and FrameNet, we have extended WordFrameNet enlarging the coverage of the original FrameNet lexicon and automatically building new local wordframenets for other languages by using the wordnets integrated in the Multilingual Central Repository (MCR) [9] (Atserias et al., 2004). We call this new resource eXtended WordFrameNet[10].

First, we have extended the coverage of FrameNet. That is, by establishing synset mappings to the FrameNet LUs, we can also add their corresponding synonyms to the frame. For instance, the frame LEADERSHIP only considers *prime_minister.n* and *premier.n*, but not *chancellor.n* which is a synonym in WordNet of those LUs. Thus, while the original FrameNet have 9,328 LUs corresponding to 6,565 synsets, eXtended WordFrameNet have 20,587 LUs. That is, more than the double. Table 4 shows the original and new LUs for the LEADERSHIP frame. In this case, 63 of the original LUs have been associated to WN synsets, thus producing 75 new LUs for this frame.

We also automatically have extended WordFrameNet to languages other than English by exploiting local wordnets aligned to the English WordNet. For instance, the Spanish synset aligned to <*prime_minister*, *premier*, *chancellor*> is <*primer_ministro*, *canciller*> and the Italian one is <*primo_ministro*>. We have alredy generated a WordFrameNet for four different languages: Spanish, Italian, Basque and Catalan. Table 3 shows the volumes of LUs for each one of these resources. Specifically, in Spanish, we obtain a WordFrameNet with 14,106 LUs. In fact, the current version of the Spanish FrameNet consists of 308 frames with 1,047 LUs[11] (Subirats and Petruck, 2003). For instance, Table 4 presents a partial view of the four versions of WordFrameNet corresponding to the LEADERSHIP

[9]http://adimen.si.ehu.es/web/MCR
[10]Available at http://adimen.si.ehu.es/web/WordFrameNet
[11]http://gemini.uab.es:9080/SFNsite/sfn-data/current-project-status

	nouns			verbs			adjectives			all		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
wn-mfs	0.75	0.75	0.75	0.64	0.64	0.64	0.80	0.80	0.80	0.69	0.69	0.69
SSI-Dijkstra	**0.84**	0.65	0.73	0.70	0.56	0.62	**0.90**	0.82	0.86	**0.78**	0.63	0.69
SSI-Dijkstra+	0.79	0.77	0.78	0.70	0.68	0.69	0.89	**0.89**	**0.89**	0.76	**0.74**	**0.75**

Table 1: Results of SSI algorithms on the *GS* dataset

frame. In this case, 96 Spanish LUs have been associated to this particular frame, while the current version of the Spanish FrameNet does not contain this frame.

XWFN	20,857
SpanishWFN	14,106
ItalianWFN	12,478
BasqueWFN	10,980
CatalanWFN	13,128

Table 3: Multilingual WFN volumes

Furthermore, we have also transported to the disambiguated LUs the knowledge currently available from other semantic resources integrated in the MCR such as SUMO (Niles and Pease, 2001), WordNet Domains (Magnini and Cavaglià, 2000), etc. For instance, now the LU corresponding to *premier.n* can also have associated the SUMO label *OccupationalRole* and its corresponding logical axioms, and the WordNet Domains *person* and *politics*. Note that when integrating multiple semantic resources such as FrameNet, WordNet and SUMO, multiple discrepancies may arise. Possibly this process can also help to improve the involved knowledge resources.

4. Conclusions and future work

We have presented an ongoing work aiming to integrate FrameNet and WordNet. The method uses a knowledge based Word Sense Disambiguation (WSD) algorithm called SSI-Dijkstra+ for assigning the appropriate synset of WordNet to the semantically related Lexical Units of a given frame from FrameNet. This algorithm relies on the use of a large knowledge base derived from WordNet and eXtended WordNet. Since the original SSI-Dijkstra requires a set of monosemous or already interpreted words, we have devised, developed and empirically tested different versions of this algorithm to deal with sets having only polysemous words. The resulting new algorithms obtain improved results over state-of-the-art. The integration of FrameNet and WordNet allows to extend the current LUs coverage. In fact, it also allows to locate conceptual areas currently uncovered by FrameNet frames. We also expect to improve the performance of the disambiguation process by using the definitions associated to the LUs. We also plan to disambiguate the Frame Elements and its corresponding definitions of a given frame. Thus, the resulting resource will also integrate the core semantic roles of FrameNet. For example, for the frame LEADERSHIP we will associate the appropriate WordNet synsets to the Frame Elements LEADER or GOVERNED. Finally, we also plan to provide WordFrameNet versions aligned to WordNet3.0 by using also the relations from the semantically annotated ”gloss corpus”.

Acknowledgement

This work has been supported by KNOW-2 (TIN2009-14715-C04-01) and KYOTO (ICT-2007-211423). Egoitz Laparra’s work is funded by a PhD grant from the Spanish government (BES-2008-001989) associated to the project KNOW (TIN2006-15049-C03-01). We want to thank the anonymous reviewers for their valuable comments.

5. References

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of GWC*, Brno, Czech Republic.

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *COLING/ACL’98*, Montreal, Canada.

A. Burchardt, K. Erk, and A. Frank. 2005. A WordNet Detour to FrameNet. In *Proceedings of the GLDV 2005 GermaNet II Workshop*, pages 408–421, Bonn, Germany.

M. Cuadros and G. Rigau. 2008. Knownet: Building a large net of knowledge from the web. In *22nd International Conference on Computational Linguistics (COLING’08)*, Manchester, UK.

J. Daudé, L. Padró, and G. Rigau. 2003. Validation and Tuning of Wordnet Mapping Techniques. In *Proceedings of RANLP*, Borovets, Bulgaria.

C. Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

FrameNet	Synset	SUMO	WordFrameNet	SpanishWFN	ItalianWFN	BasqueWFN	CatalanWFN
lead.v	*01364494-v*	Guiding	lead take direct guide	llevar conducir	portare condurre menare	eroan gidatu eraman	dirigir portar guiar conduir
command.v	*01662860-v*	Managing	command control	controlar		murriztu menderatu kontrolatu menperatu zuzendu	controlar
govern.v **rule.v**	*01763262-v*	Guiding	govern rule	gobernar regir	governare reggere amministrare dominare		governar regir
power.n	*04041746-n*	Subjective- Assesment- Attribute	power powerfulness potency	poder potencia	potenza	botere eragin	poder potència
authority.n	*04045518-n*	NormalAttribute	authority dominance say-so	autoridad dominio dominación	autorità autorevolezza balia potestà	autoritate eskumen manu aginpide aginte	autoritat domini dominació
government.n **regime.n**	*06000383-n*	Government	government regime authorities	gobierno régimen	governo autorità	gobernu erregimen botere agintari	govern règim autoritat
leader.n	*06950891-n*	Human	leader	líder dirigente autoridad	leader duce capo capintesta guru	lider buruzagi buru agintari	líder dirigent autoritat
prime_minister.n **premier.n**	*07147791-n*	OcupationRole	prime_minister premier chancellor	primer_ministro canciller	primo_ministro	lehen_ministro lehendakari kantziler	primer_ministre canceller
rector.n	*07196655-n*	OcupationRole	rector pastor minister parson curate	rector párroco vicario	rettore curato	erretore ministro bikario	rector curat vicari
chief.n **head.n**	*07311393-n*	SocialRole	chief head top_dog	responsable cabeza	responsabile capo comandante	arduradun erantzule	responsable cap director_d'escola
overlord.n	*07451003-n*	SocialRole	overlord lord master	amo señor	dominatore capo_supremo	nagusi jaun jauntxo ugazaba patroi	senyor amo
chairman.n **chairperson.n**	*07496412-n*	SocialRole	chairman chairperson chairwoman president chair	presidente moderador	presidente presidentessa	lehendakari buru presidente	president moderador
ruler.n	*07539656-n*	SocialRole	ruler	gobernador gobernante	reggitore governante	gobernatzaile gobernadore gobernari	governant governador
monarch.n	*07595596-n*	SocialRole	monarch sovereign crowned_head	monarca soberano	re regnante sovereign sovrano	monarka subirano	monarca sobirà

Table 4: Partial content of the frame LEADERSHIP in eXtended WordFrameNet

C. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, New York.

K. Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

E. Laparra and G. Rigau. 2009. Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of RANLP*, Borovets, Bulgaria.

E. Laparra, G. Rigau, and M. Cuadros. 2010. Exploring the integration of wordnet and framenet. In *Proceedings of GWC*, Mumbai, India.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proceedings of LREC*, Athens. Greece.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of ACL*, pages 280–297.

R. Mihalcea and D. Moldovan. 2001. extended wordnet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

D. Shen and M. Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the Joint Conference on (EMNLP-CoNLL)*, pages 12–21.

L. Shi and R. Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Proceedings of CICLing*, Mexico.

C. Subirats and M. Petruck. 2003. Surprise: Spanish framenet! In *Proceedings of the International Congress of Linguists*, Praga.

S. Tonelli and E. Pianta. 2009. A novel approach to mapping framenet lexical units to wordnet synsets. In *Proceedings of IWCS-8*, Tilburg, The Netherlands.