Aligning FrameNet and WordNet based on Semantic Neighborhoods

Óscar Ferrández[1], **Michael Ellsworth**[2], **Rafael Muñoz**[1], **Collin F. Baker**[2]

(1) Department of Software and Computing Systems, University of Alicante, Spain
(2) International Computer Science Institute (ICSI), Berkeley, CA, USA
ofe@dlsi.ua.es, infinity@icsi.berkeley.edu, rafael@dlsi.ua.es, collinb@icsi.berkeley.edu

Abstract

This paper presents an algorithm for aligning FrameNet lexical units to WordNet synsets. Both, FrameNet and WordNet, are well-known as well as widely-used resources by the entire research community. They help systems in the comprehension of the semantics of texts, and therefore, finding strategies to link FrameNet and WordNet involves challenges related to a better understanding of the human language. Such deep analysis is exploited by researchers to improve the performance of their applications. The alignment is achieved by exploiting the particular characteristics of each lexical-semantic resource, with special emphasis on the explicit, formal semantic relations in each. Semantic neighborhoods are computed for each alignment of lemmas, and the algorithm calculates correlation scores by comparing such neighborhoods. The results suggest that the proposed algorithm is appropriate for aligning the FrameNet and WordNet hierarchies. Furthermore, the algorithm can aid research on increasing the coverage of FrameNet, building FrameNets in other languages, and creating a system for querying a joint FrameNet-WordNet hierarchy.

1. Introduction

Human Language Technologies attempts to simulate human linguistic behavior, however there is no consensus on how closely it should be simulated. In such a complex task, the computer has to be aware of the structures used by the language as well as the global discourse. Obviously, the comprehension of the semantics is of paramount importance to build such machines. This is the reason why the research community is keenly interested in the construction of semantic resources capable of modelling the semantics of texts. Moreover, researches have often used these resources to improve the performance of their applications. Two of these lexical-semantic resources are FrameNet (Fontenelle, 2003; Ruppenhofer et al., 2006) and WordNet (Miller et al., 1990).

The FRAMENET project[1] is creating a highly detailed lexicon of English based on the theory of frame semantics, and supported by corpus evidence. The goal is to document the range of semantic and syntactic combinatory possibilities (namely *valences*) of each word sense (called *lexical unit, LU*) by examples from attestations taken from naturalistic corpora, mainly from the British National Corpus (BNC)[2], rather than constructed by a linguist or lexicographer.

To understand the semantic network that represents FrameNet it is necessary to describe the main concepts encoded in it. An `LU` is the pairing of a word with a meaning. Typically, each sense of a polysemous word belongs to a different semantic `frame`, which is a script-like conceptual structure that describes a particular type of situation, object, or event and the participants and props involved therein (called `frame elements, FEs`). Moreover, FN also provides frame-to-frame relations. Each relation represents an asymmetric link between two frames, also connecting the FEs that participate in each particular relation.

The major product of this work, the FrameNet lexical database (available upon request)[3], currently contains more than 11,500 lexical units (more than 6,800 of which are fully annotated) in more than 950 semantic frames, exemplified in more than 150,000 annotated sentences. It has gone through five releases, is freely available for research purposes, and is now in use by hundreds of researchers, teachers, and students around the world. Active research projects are now seeking to produce comparable frame-semantic lexicons for other languages (e.g. Spanish FrameNet[4] (Subirats and Petruck, 2003)) and to devise a means of automatically labelling running text with semantic frame information (Erk and Padó, 2006; Johansson and Nugues, 2007a).

WORDNET (Miller et al., 1990), an electronic lexical database of English, is considered to be one of the most important resources available to researchers in computational linguistics, text analysis, and several other related areas. Its design is inspired by psycholinguistic and computational theories of human lexical memory. English nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each representing one underlying lexicalized concept. Synsets are interlinked by means of conceptual-semantic and lexical relations, obtaining as a result a lexico-semantic interlinked concept network of more than 200,000 word-sense pairs.

This paper presents an alignment algorithm of FrameNet and WordNet. Both are well-known as well as widely used lexical-semantic resources. With this alignment, we will link each Lexical Unit (LU) of FrameNet to a specific WordNet synset. Achieving the complete alignment of the whole set of LUs will allow us to build a joint hierarchy, which may help to remedy some of the weaknesses of both resources. Then, measures which were previously derived from only one of them will be able to be applied to the joint hierarchy. Additionally, it will also allow us to extend the coverage of FrameNet via the WordNet relations. For in-

[1] `http://framenet.icsi.berkeley.edu`
[2] `http://www.natcorp.ox.ac.uk/`
[3] `http://framenet.icsi.berkeley.edu`
[4] `http://gemini.uab.es/SFN`

stance, if there are WordNet synonyms and/or hyponyms of a synset that has been aligned to a specific LU, these synonyms and/or hyponyms could be considered as LUs that evoke the frame corresponding to the alignment. Furthermore, the alignment could also help in the construction of FrameNets in other languages besides English by linking English frames to synsets of WordNets in other languages.

2. Related Work

Finding strategies to link FrameNet to WordNet and evaluating the alignments in applications involve challenges related to algorithms for word sense disambiguation, measuring textual similarity, etc. The last few years have seen a surge of interest in modeling and designing approaches aimed at achieving these goals.

For instance, in (Pennacchiotti et al., 2008) the authors proposed two models for automatic induction of FrameNet Lexical Units: (1) a *distributional model*, which induces new lexical units by modelling existing frames and unknown lexical units in the form of co-occurrence vectors computed over a corpus; and (2) a *WordNet-based model* based on the intuition that senses able to evoke a frame can be detected by jointly considering the WordNet synsets activated by all LUs of the frame. With these techniques, they extend FrameNet by discovering new lexical units, to some extent following the line or research proposed at SemEval-2007 (Baker et al., 2007) for using learning approaches to automatically annotate unseen frames and roles.

Johansson and Nugues (2007b) also propose an approach to add new units to the FrameNet lexical database. They deal with the sparsity of FrameNet by using the WordNet-based Lesk measure and the WordNet hypernym hierarchy as feature vectors for a machine learning classifier. For each frame they train a classifier that will determine whether a lemma belongs to the frame or not.

Burchardt et al. (2005) extend FrameNet via WordNet by means of a rule-based system based on a word sense disambiguation tool to ascertain the correct synset and the extraction of WordNet relatives (i.e. synonyms, hypernyms and antonyms) together with the candidate frames. The authors take this approach, which they call "a detour via WordNet", in order to cope with sparse data in the automatic frame assignment task, which requires annotation of previously unseen lexical units.

Cao et al. (2008) implemented a similarity function exploiting different WordNet information in order to obtain a set of WordNet senses able to evoke the same frame. Their approach was applied to LU induction for the English FrameNet as well as for Italian FrameNet via MultiWordNet.

Laparra and Rigau (2009) describe a knowledge-based word sense disambiguation algorithm that uses a large-scale knowledge base derived from WordNet, and partially integrates FrameNet and WordNet to extend FrameNet coverage. With this extension they expect: (i) to enrich WordNet with frame semantic information; and possibly (ii) to extend FrameNet to other languages other than English.

Tonelli and Pianta (2009) and Tonelli and Pighin (2009) proposed a method to map FrameNet LUs to WordNet synsets by computing a similarity measure between LU definitions and WordNet glosses, enriched with other features such as a synset-frame overlap and WordNet domains. The main aim of Tonelli's work is to automatically increase the number of LUs for each frame by importing all synonyms from the mapped synset.

Shi and Mihalcea (2005) presented a semi-automatic approach that uses VerbNet as a bridge between FrameNet and WordNet to align verbs, and this mapping is later part of a system that assigns frames to verbs in open text (Honnibal and Hawker, 2005).

In (Ide, 2006), the author described an approach that maps verbs associated with FrameNet to appropriate WordNet 2.0 senses, demonstrating its applicability to additional tasks such as assignment of WordNet senses to word lists used in attitude and opinion analysis, and collapsing WordNet senses into coarse-grained groupings.

Finally, Chow and Webster (2006) propose a statistical distribution approach to reuse and integrate information from SUMO[5], WordNet, and FrameNet. The mapping is accomplished through the verbs covered by both FrameNet and WordNet, taking the shared lexical knowledge as learning data to map SUMO concepts with FrameNet frames. They used the verb mapping presented in (Shi and Mihalcea, 2005) and the motivation of the FrameNet-SUMO mapping achieved consisted of the construction of a comprehensive knowledge base for Semantic Role Labeling (SLR).

3. The Alignment

The fundamental idea is to exploit the relations that occur in each lexical resource in order to achieve correlation scores between the synsets that contain a specific $lemma.pos$ ("synsets containing a lemma", SSL) and the LUs that also contain the same $lemma.pos$. Program 1 details the algorithm step by step.

The algorithm starts with a particular *lemma.pos*, perhaps occurring in several SSLs and also in several LUs. It looks up all of the senses associated with that *lemma.pos*, traversing each of the relations in the resources in turn to construct one WordNet neighborhood and one FrameNet neighborhood with the starting word sense at the center. Then, for each neighbor (appearing in any or in both neighborhoods): we obtain its distance with regards to the centering word of each neighborhood (in the event this neighbor does not exist we will get a MAX_VALUE); we subtract these distances to get small numbers if they are similar; and, we take the inverse to produce the similarity score. All neighbors similarity scores are accumulated, multiplied by the weight assigned to the current FrameNet-WordNet relation pair and normalized by the total amount of processed neighbors. To establish the relevance of each FrameNet and WordNet relation, we use a set of 100 manually-created LU-SSL alignments as a seed for acquiring further alignments.

Additionally, following previous research work (especially (Tonelli and Pianta, 2009)), we decided to add a factor that improves the alignment by calculating the degree of textual similarity between the LU definition and the WordNet gloss. Finally, a variety of methods could be used to

[5]The Suggested Upper Merged Ontology (SUMO), `http://www.ontologyportal.org/`.

make the alignment decision, but the most obvious is perhaps best-first. We use this in our experiments, counting as correct those alignments which obtain the best score for each LU-SSL pair.

Figure 1 depicts the formula that represents the inner steps of the algorithm, where C is the correlation between an LU and an SSL, S; R_{FN} is a frame-to-frame relation or lemma-to-frame relation type; R_{WN} is a synset-to-synset or word-to-synset relation type; W is a function which weights the expected informativeness of each pair of WordNet-FrameNet relation types; λ is a word in the vicinity of LU (along relation R_{FN}) and/or S (along relation R_{WN}); $\|\lambda\|$ is the number of words in the WordNet and FrameNet neighborhoods along the relevant relations; $d_{R_{FN}}$ is the LU-to-LU distance function from traversing the FrameNet hierarchy along relation R_{FN}; $d_{R_{WN}}$ is the sense-to-sense distance function from traversing the WordNet hierarchy along relation R_{WN}; and α is a small constant (currently set to 1) to prevent division by zero as well as to prevent complete swamping by good individual correlations.

Regarding R_{FN} and R_{WN}, each FN/WN relation is associated with a score; these are shown in Table 1. For FrameNet we distinguish whether the relation goes up ($PARENT$) or down ($CHILD$) in the hierarchy. All scores were obtained heuristically by considering the meaningfulness of each relation, and by training experiments on a manually-created dataset of 100 LU-SSL alignments[6]. The work presented in (Moldovan and Novischi, 2002) served as inspiration for setting the weights for R_{WN}s. Note that the most relevant FrameNet relations have high values in comparison with the WordNet relations; this is due to the fact that the neighborhood sizes of the two resources are not balanced (the WordNet neighborhoods are much bigger), so the FrameNet scores have been set higher to compensate for the difference in size.

FrameNet Relation	PARENT	CHILD
Membership	10 (points)	
Inheritance	9	5
Perspective on	1	0.8
SubFrame	0.8	0.6
Precedes	0.7	0.7
Causative	0.5	0.4
Inchoative	0.5	0.4
Using	0.2	0.1
See also	0.8	0.7

Table 1: FrameNet relations scores.

In order to obtain the neighbor distances (i.e. $d_{R_{FN}}$ and $d_{R_{WN}}$), the relation scores are multiplied by the depth of the neighbor measured from the starting LU or SSL. In the current version of the algorithm, the neighborhood consists of all the neighbors connected by a single relation to the starting $lemma.pos$.

With regard to the FrameNet-WordNet relation pair (i.e. $W(R_{FN}, R_{WN})$), at the current state of this research all

[6]This manual alignment can be downloaded at `http://www.dlsi.ua.es/~ofe/berkeley/`.

Relation	Score
Synonymy	1
Hypernymy	0.8
Hyponymy	0.7
Antonymy	0.1
Entailment	0.7
Cause to	0.5
Derived forms	0.7
Holonomy	0.5
Meronomy	0.5
Attributes	0.5
Coordinate terms	0.5

Table 2: WordNet relations scores.

W-values are set to one, so that the final alignment is determined by the other values within the formula. Our highest-priority future work will consist of establishing these values properly. Finally, after checking several string-text similarity algorithms, we chose the traditional n-dimensional Euclidean distance to measure the similarity between the LU definition and the WordNet gloss.

To measure the similarity of the strings, we set n as the number of distinct items (characters) that occur in a pair of stems (one from the LU definition and the other from the gloss); and each Euclidean point represents the number of times that each item appear in each stem respectively. Then the distance is computed for each stem of one text regarding all stems of the other text, and the maximum value is returned. Afterwards, to obtain a normalized similarity score, these values are summed and divided by the number of the stems' combinations processed, which will also correspond with the length of the shortest text.

4. Resource Evaluation

Our main target is to obtain a mapping between the LUs encoded in FrameNet 1.3 to the synsets of WordNet 3.0, and evaluate its accuracy. To achieve this, we extracted the entire set of LUs (i.e. 10,195 LUs), and we tried to find the WordNet counterparts of the *lemma.poses* involved in these LUs. Unfortunately, we were not able to find a counterpart in WordNet 3.0 for every LU mostly due to distinct part-of-speeches in both resources. Finally, after resolving meaningless differences such as hyphenation (e.g. *tip-top.adj* in FrameNet vs. *tiptop.adj* in WordNet), we created an evaluation framework made up of 9,612 LUs, and for alignment purposes, we ran the algorithm over the set of 7,830 distinct *lemma.pos* contained in these 9,612 LUs. The final alignments can be downloaded in XML format at `http://www.dlsi.us.es/~ofe/berkeley/`.

In order to assess the accuracy of our alignment procedure, we used the gold standard provided by Sara Tonelli (Tonelli and Pianta, 2009), consisting of a set of manually annotated mappings between LUs and WordNet 1.6 synsets. To use this gold-standard we had to map each WordNet 1.6 synset to its corresponding WordNet 3.0 synset, and for consistency with our framework and algorithm, we also had to discard those mappings that have different part-of-

Program 1 The alignment algorithm in detail.

```
INPUT: lemma+PoS FrameNet LUs related to this lemma+PoS: LUs = {LU1, ..., LUn}
WordNet SSLs related to this lemma+PoS: SSL = {SSL1, ..., SSLn}

 for each LU and SSL of the lemma+PoS
  for each FrameNet-relation and WordNet-relation
   - obtain the lemma+PoS FrameNet neighborhood traversing the given relation
   - obtain the lemma+PoS WordNet neighborhood traversing the given relation
   for each neighbor in the neighborhood of the LU or SSL
    - calculate the distance between the neighbor and starting LU
    - calculate the distance between the neighbor and starting SSL
      (if there is no such neighbor, use a default maximum)
    - subtract these distances
      (when the distances are similar, we get a small number)
    - take the inverse to produce a similarity score
      (now good correlation gives a big number)
    - aggregate the score for the neighborhood (by summing)
   - normalize by dividing by the number of neighborhood lemmas
   - multiply by a weight for the current WordNet/FrameNet relation pair
   - aggregate the score per relation (by summing)
  - obtain the textual similarity degree between the LU-def & WN-gloss
  - aggregate the similarity score by summing
  - judge whether the correlation for the mapping is good enough
    (the best-scoring pair SSL-LU will be matched)
  - if so, join the LU and SSL in the joint hierarchy
- if not, move on
```

$$C(LU,S) = \sum_{R_{FN}} \sum_{R_{WN}} \left(\frac{W(R_{FN},R_{WN})}{\|\lambda\|} \sum_{\lambda} \frac{1}{|d_{R_{FN}}(LU,\lambda) - d_{R_{WN}}(S,\lambda)| + \alpha(R_{FN},R_{WN})} \right) + TextSim(LU_{def}, S_{gloss})$$

Figure 1: The correlation score formula.

speech or lemmata. Moreover, to make the gold-standard bigger, we also included in the evaluation the WordNet-FrameNet Verbal Mapping from (Shi and Mihalcea, 2005). This dataset includes a mapping between FrameNet 1.2 verb LUs and WordNet 2.0 verbal senses. To be able to evaluate our alignment using this mapping, we had to translate WordNet 2.0 sense keys to version 3.0 as well as check which FrameNet 1.2 LUs involved in the mapping also appear in FrameNet 1.3.

As a result, the final gold-standard mapping contains a total of 2,874 verbs, 124 nouns, 57 adjectives and 1 adverb.

Table 3 shows the overall accuracy reached by our algorithm as well as individual rates by each part-of-speech from the gold standard. The table also gives the most frequent sense (MFS) as a simple benchmark for assessing the performance of our alignments. The MFS approach computes the alignments according to the WordNet sense ranking, always tagging as the correct alignment the first WordNet sense of the synset.

Regarding the monosemous entries of the gold standard, 22% of the total amount of entries are monosemous. The distribution of such entries by part-of-speech is interesting: the relative polysemy of verbs and nouns is 78% vs. 84% respectively. Consequently, due to the fact thet the nouns set has more polysemous entries, the behavior of our algorithm obtains a more significant improvement (over the MFS benchmark) when dealing with verbs than nouns.

5. Conclusions and Future Work

This paper proposes an automatic alignment of FrameNet and WordNet that exploits the particular idiosyncrasies of each hierarchy, obtaining correlation scores between FrameNet LUs and WordNet synsets. Results reveals that the algorithm is appropriate for aligning these two resources, obtaining an overall accuracy greater than 77%.

With our algorithm, we leave open a line for further research on issues related to the use of the alignments. For instance, once FrameNet coverage can be extended through WordNet relations, inferences which were developed for one of the two resources/hierarchies can be applied to the joint hierarchy, and these alignments can assist in the construction of FrameNets in other languages.

In the next stages of this research, we would like to work on: (i) heuristically setting the importance of each FrameNet-WordNet relation pair within the alignment process; (ii) expanding the gold-standard dataset by manually annotating more examples; (iii) using this extended gold-standard in order to learn and better set the value of those parameters of the algorithm which can be trained (i.e. W, the weight in the distance functions per relation, α and suitable neighborhood sizes per relation); (iv) establishing a criterion for the minimum value which must be obtained in order to consider an alignment correct (rather than using the best-first choice), which would allow us to obtain results in terms of precision, recall and F-measure; and (v) integrat-

Our Alignment				MFS			
overall Accuracy				overall Accuracy			
0.772				0.657			
Nouns	Verbs	Adjs	Advs	Nouns	Verbs	Adjs	Advs
0.78	0.77	0.81	1	0.70	0.58	0.80	1

Table 3: Overall and grammatical-class dependent results.

ing a machine learning classifier trained on our alignment features, with the aim of improving the final alignment performance.

Acknowledgments

This research has been funded by the Spanish Visiting Senior Scientific program at the International Computer Science Institute (Berkeley, CA), by Generalitat Valenciana GV PROMETEO/2009/119 and by the Alicante University post-doctoral fellowship program funded by *Fundación CajaMurcia*.

6. References

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June. Association for Computational Linguistics.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A wordnet detour to framenet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining word sense and usage for modeling frame semantics. In *Proceedings of STEP 2008*, Venice, Italy.

Ian C. Chow and Jonathan J. Webster. 2006. Mapping framenet and sumo with wordnet verb: Statistical distribution of lexical-ontological realization. *Mexican International Conference on Artificial Intelligence*, 0:262–268.

Katrin Erk and Sebastian Padó. 2006. Shalmaneser – A Flexible Toolbox For Semantic Role Assignment. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC-2006)*.

Thierry Fontenelle, editor. 2003. *International Journal of Lexicography–Special Issue on FrameNet*, volume 16. Oxford University Press.

Matthew Honnibal and Tobias Hawker. 2005. Identifying framenet frames for verbs from a real-text corpus. In *Australasian Language Technology Workshop*.

Nancy Ide. 2006. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Alexander F. Gelbukh, editor, *CICLing*, volume 3878 of *Lecture Notes in Computer Science*, pages 13–27. Springer.

Richard Johansson and Pierre Nugues. 2007a. Lth: Semantic structure extraction using nonprojective dependency trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230, Prague, Czech Republic, June. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2007b. Using WordNet to Extend FrameNet Coverage. In *FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages, NOVALIDA 2007*, Tartu, Estonia, May.

Egoitz Laparra and German Rigau. 2009. Integrating WordNet and FrameNet using a knowledge-based Word Sense Disambiguation algorithm. In *Proceedings of Recent Advances in Natural Language Processing (RANLP09)*, Borovets, Bulgaria, September.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Dan Moldovan and Adrian Novischi. 2002. Lexical chains for question answering. In *Proceedings of COLING 2002*.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii, October.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico*, pages 100–111.

Carlos Subirats and Miriam R. L. Petruck. 2003. Surprise: Spanish FrameNet! In *Proceedings of the International Congress of Linguists*, Prague, Czech Republic.

Sara Tonelli and Emanuele Pianta. 2009. A novel approach to mapping FrameNet lexical units to WordNet synsets. In *Proceedings of IWCS-8*, Tilburg, The Netherlands.

Sara Tonelli and Daniele Pighin. 2009. New features for framenet - wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 219–227, Boulder, Colorado, June. Association for Computational Linguistics.