Scaling up Automatic Cross-Lingual Semantic Role Annotation

Lonneke van der Plas
Department of Linguistics
University of Geneva
Geneva, Switzerland

Paola Merlo
Department of Linguistics
University of Geneva
Geneva, Switzerland

James Henderson
Department of Computer Science
University of Geneva
Geneva, Switzerland

`{Lonneke.vanderPlas,Paola.Merlo,James.Henderson}@unige.ch`

Abstract

Broad-coverage semantic annotations for training statistical learners are only available for a handful of languages. Previous approaches to cross-lingual transfer of semantic annotations have addressed this problem with encouraging results on a small scale. In this paper, we scale up previous efforts by using an automatic approach to semantic annotation that does not rely on a semantic ontology for the target language. Moreover, we improve the quality of the transferred semantic annotations by using a joint syntactic-semantic parser that learns the correlations between syntax and semantics of the target language and smooths out the errors from automatic transfer. We reach a labelled F-measure for predicates and arguments of only 4% and 9% points, respectively, lower than the upper bound from manual annotations.

1 Introduction

As data-driven techniques tackle more and more complex natural language processing tasks, it becomes increasingly unfeasible to use complete, accurate, hand-annotated data on a large scale for training models in all languages. One approach to addressing this problem is to develop methods that automatically generate annotated data by transferring annotations in parallel corpora from languages for which this information is available to languages for which these data are not available (Yarowsky et al., 2001; Fung et al., 2007; Padó and Lapata, 2009).

Previous work on the cross-lingual transfer of semantic annotations (Padó, 2007; Basili et al., 2009) has produced annotations of good quality for test sets that were carefully selected based on semantic ontologies on the source and target side. It has been suggested that these annotations could be used to train semantic role labellers (Basili et al., 2009).

In this paper, we generate high-quality broad-coverage semantic annotations using an automatic approach that does not rely on a semantic ontology for the target language. Furthermore, to our knowledge, we report the first results on using joint syntactic-semantic learning to improve the quality of the semantic annotations from automatic cross-lingual transfer. Results on correlations between syntax and semantics found in previous work (Merlo and van der Plas, 2009; Lang and Lapata, 2010) have led us to make use of the available syntactic annotations on the target language. We use the semantic annotations resulting from cross-lingual transfer combined with syntactic annotations to train a joint syntactic-semantic parser for the target language, which, in turn, re-annotates the corpus (See Figure 1). We show that the semantic annotations produced by this parser are of higher quality than the data on which it was trained.

Given our goal of producing broad-coverage annotations in a setting based on an aligned corpus, our choices of formal representation and of labelling scheme differ from previous work (Padó, 2007; Basili et al., 2009). We choose a dependency representation both for the syntax and semantics because relations are expressed as direct arcs between words. This representation allows cross-lingual transfer to use word-based alignments directly, eschewing the need for complex constituent-alignment algorithms.
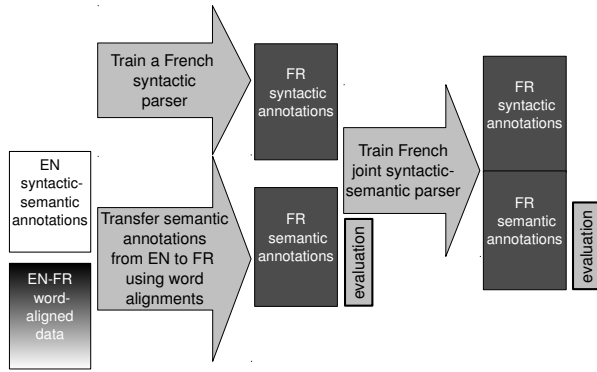
299

Figure 1: System overview

We choose the semantic annotation scheme defined by Propbank, because it has broad coverage and includes an annotated corpus, contrary to other available resources such as FrameNet (Fillmore et al., 2003) and is the preferred annotation scheme for a joint syntactic-semantic setting (Merlo and van der Plas, 2009). Furthermore, Monachesi et al. (2007) showed that the PropBank annotation scheme can be used for languages other than English directly.

2 Cross-lingual semantic transfer

Data-driven induction of semantic annotation based on parallel corpora is a well-defined and feasible task, and it has been argued to be particularly suitable to semantic role label annotation because cross-lingual parallelism improves as one moves to more abstract linguistic levels of representation. While Hwa et al. (2002; 2005) find that direct syntactic dependency parallelism between English and Spanish concerns 37% of dependency links, Padó (2007) reports an upper-bound mapping correspondence calculated on gold data of 88% F-measure for individual semantic roles, and 69% F-measure for whole scenario-like semantic frames. Recently, Wu and Fung (2009a; 2009b) also show that semantic roles help in statistical machine translation, capitalising on a study of the correspondence between English and Chinese which indicates that 84% of roles transfer directly, for PropBank-style annotations. These results indicate high correspondence across languages at a shallow semantic level.

Based on these results, our transfer of semantic annotations from English sentences to their French translations is based on a very strong mapping hypothesis, adapted from the Direct Correspondence Assumption for syntactic dependency trees by Hwa et al. (2005).

> **Direct Semantic Transfer (DST)** For any pair of sentences E and F that are translations of each other, we transfer the semantic relationship $R(x_E, y_E)$ to $R(x_F, y_F)$ if and only if there exists a word-alignment between x_E and x_F and between y_E and y_F, and we transfer the semantic property $P(x_E)$ to $P(x_F)$ if and only if there exists a word-alignment between x_E and x_F.

The relationships which we transfer are semantic role dependencies and the properties are predicate senses. We introduce one constraint to the direct semantic transfer. Because the semantic annotations in the target language are limited to verbal predicates, we only transfer predicates to words the syntactic parser has tagged as a verb.

As reported by Hwa et al. (2005), the direct correspondence assumption is a strong hypothesis that is useful to trigger a projection process, but will not work correctly for several cases.

We used a filter to remove obviously incomplete annotations. We know from the annotation guidelines used to annotate the French gold sentences that all verbs, except modals and realisations of the verb *être*, should receive a predicate label. We define a filter that removes sentences with missing predicate labels based on PoS-information in the French sentence.

2.1 Learning joint syntactic-semantic structures

We know from previous work that there is a strong correlation between syntax and semantics (Merlo and van der Plas, 2009), and that this correlation has been successfully applied for the unsupervised induction of semantic roles (Lang and Lapata, 2010). However, previous work in machine translation leads us to believe that transferring the correlations between syntax and semantics across languages would be problematic due to argument-structure divergences (Dorr, 1994). For example, the English verb *like* and the French verb *plaire* do not share correlations between syntax and semantics. The verb *like* takes an A0 subject and an A1

direct object, whereas the verb *plaire* licences an A1 subject and an A0 indirect object.

We therefore transfer semantic roles cross-lingually based only on lexical alignments and add syntactic information after transfer. In Figure 1, we see that cross-lingual transfer takes place at the semantic level, a level that is more abstract and known to port relatively well across languages, while the correlations with syntax, that are known to diverge cross-lingually, are learnt on the target language only. We train a joint syntactic-semantic parser on the combination of the two linguistic levels that learns the correlations between these structures in the target language and is able to smooth out errors from automatic transfer.

3 Experiments

We used two statistical parsers in our transfer of semantic annotations from English to French, one for syntactic parsing and one for joint syntactic-semantic parsing. In addition, we used several corpora.

3.1 The statistical parsers

For our syntactic-semantic parsing model, we use a freely-available parser (Henderson et al., 2008; Titov et al., 2009). The probabilistic model is a joint generative model of syntactic and semantic dependencies that maximises the joint probability of the syntactic and semantic dependencies, while building two separate structures.

For the French syntactic parser, we used the dependency parser described in Titov and Henderson (2007). We train the parser on the dependency version of the French Paris treebank (Candito et al., 2009), achieving 87.2% labelled accuracy on this data set.

3.2 Data

To transfer semantic annotation from English to French, we used the Europarl corpus (Koehn, 2003)[1]. We word-align the English sentences to the French sentences automatically using GIZA++ (Och

and Ney, 2003) and include only intersective alignments. Furthermore, because translation shifts are known to pose problems for the automatic projection of semantic roles across languages (Padó, 2007), we select only those parallel sentences in Europarl that are direct translations from English to French, or vice versa. In the end, we have a word-aligned parallel corpus of 276-thousand sentence pairs.

Syntactic annotation is available for French. The French Treebank (Abeillé et al., 2003) is a treebank of 21,564 sentences annotated with constituency annotation. We use the automatic dependency conversion of the French Treebank into dependency format provided to us by Candito and Crabbé and described in Candito et al. (2009).

The Penn Treebank corpus (Marcus et al., 1993) merged with PropBank labels (Palmer et al., 2005) and NomBank labels (Meyers, 2007) is used to train the syntactic-semantic parser described in Subsection 3.1 to annotate the English part of the parallel corpus.

3.3 Test sets

For testing, we used the hand-annotated data described in (van der Plas et al., 2010). One-thousand French sentences are extracted randomly from our parallel corpus without any constraints on the semantic parallelism of the sentences, unlike much previous work. We randomly split those 1000 sentences into test and development set containing 500 sentences each.

4 Results

We evaluate our methods for automatic annotation generation twice: once after the transfer step, and once after joint syntactic-semantic learning. The comparison of these two steps will tell us whether the joint syntactic-semantic parser is able to improve semantic annotations by learning from the syntactic annotations available. We evaluate the models on unrestricted test sets[2] to determine if our methods scale up.

Table 1 shows the results of automatically annotating French sentences with semantic role annotation. The first set of columns of results re-

---

[1] As is usual practice in preprocessing for automatic alignment, the datasets were tokenised and lowercased and only sentence pairs corresponding to a one-to-one sentence alignment with lengths ranging from one to 40 tokens on both French and English sides were considered.

[2] Due to filtering, the test set for the *transfer (filter)* model is smaller and not directly comparable to the other three models.

		Predicates						Arguments (given predicate)					
		Labelled			Unlabelled			Labelled			Unlabelled		
		Prec	Rec	F	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
1	Transfer (no filter)	50	31	38	91	55	69	61	48	54	72	57	64
2	Transfer (filter)	51	46	**49**	92	84	**88**	65	51	**57**	76	59	**67**
3	Transfer+parsing (no filter)	71	29	42	97	40	57	77	57	**65**	87	64	**74**
4	Transfer+parsing (filter)	61	50	**55**	95	78	**85**	71	52	60	83	61	70
5	Inter-annotator agreement	61	57	59	97	89	93	73	75	74	88	91	89

Table 1: Percent recall, precision, and F-measure for predicates and for arguments given the predicate, for the four automatic annotation models and the manual annotation.

ports labelling and identification of predicates and the second set of columns reports labelling and identification of arguments, respectively, for the predicates that are identified. The first two rows show the results when applying direct semantic transfer. Rows three and four show results when using the joint syntactic-semantic parser to re-annotate the sentences. For both annotation models we show results when using the filter described in Section 2 and without the filter.

The most striking result that we can read from Table 1 is that the joint syntactic-semantic learning step results in large improvements, especially for argument labelling, where the F-measure increases from 54% to 65% for the unfiltered data. The parser is able to outperform the quality of the semantic data on which it was trained by using the information contained in the syntax. This result is in accordance with results reported in Merlo and Van der Plas (2009) and Lang and Lapata (2010), where the authors find a high correlation between syntactic functions and PropBank semantic roles.

Filtering improves the quality of the transferred annotations. However, when training a parser on the annotations we see that filtering only results in better recall scores for predicate labelling. This is not surprising given that the filters apply to completeness in predicate labelling specifically. The improvements from joint syntactic-semantic learning for argument labelling are largest for the unfiltered setting, because the parser has access to larger amounts of data. The filter removes 61% of the data.

As an upper bound we take the inter-annotator agreement for manual annotation on a random set of 100 sentences (van der Plas et al., 2010), given in the last row of Table 1. The parser reaches an F-measure on predicate labelling of 55% when using filtered data, which is very close to the upper bound (59%). The upper bound for argument inter-annotator agreement is an F-measure of 74%. The parser trained on unfiltered data reaches an F-measure of 65%. These results on unrestricted test sets and their comparison to manual annotation show that we are able to scale up cross-lingual semantic role annotation.

5 Discussion and error analysis

A more detailed analysis of the distribution of improvements over the types of roles further strengthens the conclusion that the parser learns the correlations between syntax and semantics. It is a well-known fact that there exists a strong correlation between syntactic function and semantic role for the A0 and A1 arguments: A0s are commonly mapped onto subjects and A1s are often realised as direct objects (Lang and Lapata, 2010). It is therefore not surprising that the F-measure on these types of arguments increases by 12% and 15%, respectively, after joint-syntactic semantic learning. Since these arguments make up 65% of the roles, this introduces a large improvement. In addition, we find improvements of more than 10% on the following adjuncts: AM-CAU, AM-LOC, AM-MNR, and AM-MOD that together comprise 9% of the data.

With respect to predicate labelling, comparison of the output after transfer with the output after parsing (on the development set) shows how the parser smooths out transfer errors and how interlingual divergences can be solved by making use of the variations we find intra-lingually. An example is given in Figure 2. The first line shows the predicate-argument structure given by the English

EN (source)	Postal [$_{A1}$ services] [$_{AM\text{-}MOD}$ must] [$_{CONTINUE.01}$ continue] [$_{C\text{-}A1}$ to] be public services.
FR (transfer)	Les [$_{A1}$services] postaux [$_{AM\text{-}MOD}$ doivent] [$_{CONTINUE.01}$rester] des services publics.
FR (parsed)	Les [$_{A1}$ services] postaux [$_{AM\text{-}MOD}$ doivent] [$_{REMAIN.01}$rester] des [$_{A3}$ services] publics.

Figure 2: Differences in predicate-argument labelling after transfer and after parsing

syntactic-semantic parser to the English sentence. The second line shows the French translation and the predicate-argument structure as it is transferred cross-lingually following the method described in Section 2. Transfer maps the English predicate label CONTINUE.01 onto the French verb *rester*, because these two verbs are aligned. The first occurrence of *services* is aligned to the first occurrence of *services* in the English sentence and gets the A1 label. The second occurrence of *services* gets no argument label, because there is no alignment between the C-A1 argument *to*, the head of the infinitival clause, and the French word *services*. The third line shows the analysis resulting from the syntactic-semantic parser that has been trained on a corpus of French sentences labelled with automatically transferred annotations and syntactic annotations. The parser has access to several labelled examples of the predicate-argument structure of *rester*, which in many other cases is translated with *remain* and has the same predicate-argument structure as *rester*. Consequently, the parser re-labels the verb with REMAIN.01 and labels the argument with A3.

Because the languages and annotation framework adopted in previous work are not directly comparable to ours, and their methods have been evaluated on restricted test sets, results are not strictly comparable. But for completeness, recall that our best result for predicate identification is an F-measure of 55% accompanied with an F-measure of 60% for argument labelling. Padó (2007) reports a 56% F-measure on transferring FrameNet roles, knowing the predicate, from an automatically parsed and semantically annotated English corpus. Padó and Pitel (2007), transferring semantic annotation to French, report a best result of 57% F-measure for argument labelling given the predicate. Basili et al. (2009), in an approach based on phrase-based machine translation to transfer FrameNet-like annotation from English to Italian, report 42% recall in identifying predicates and an aggregated 73% recall of identifying predicates and roles given these pred-

icates. They do not report an unaggregated number that can be compared to our 60% argument labelling. In a recent paper, Annesi and Basili (2010) improve the results from Basili et al. (2009) by 11% using Hidden Markov Models to support the automatic semantic transfer. Johansson and Nugues (2006) trained a FrameNet-based semantic role labeller for Swedish on annotations transferred cross-lingually from English parallel data. They report 55% F-measure for argument labelling given the frame on 150 translated example sentences.

6 Conclusions

In this paper, we have scaled up previous efforts of annotation by using an automatic approach to semantic annotation transfer in combination with a joint syntactic-semantic parsing architecture. We propose a direct transfer method that requires neither manual intervention nor a semantic ontology for the target language. This method leads to semantically annotated data of sufficient quality to train a syntactic-semantic parser that further improves the quality of the semantic annotation by joint learning of syntactic-semantic structures on the target language. The labelled F-measure of the resulting annotations for predicates is only 4% point lower than the upper bound and the resulting annotations for arguments only 9%.

Acknowledgements

References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a treebank for French. In *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers.

P. Annesi and R. Basili. 2010. Cross-lingual alignment of FrameNet annotations through Hidden Markov Models. In *Proceedings of CICLing*.

R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti, 2009. *Computational Linguistics and Intelligent Text Processing*, chapter Cross-Language Frame Semantics Transfer in Bilingual Corpora, pages 332–345. Springer Berlin / Heidelberg.

M.-H. Candito, B. Crabbé, P. Denis, and F. Guérin. 2009. Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

B. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

C. J. Fillmore, R. Johnson, and M.R.L. Petruck. 2003. Background to FrameNet. *International journal of lexicography*, 16.3:235–250.

P. Fung, Z. Wu, Y. Yang, and D. Wu. 2007. Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*.

J. Henderson, P. Merlo, G. Musillo, and I. Titov. 2008. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *Proceedings of CONLL 2008*, pages 178–182.

R. Hwa, P. Resnik, A. Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*.

R. Hwa, P. Resnik, A.Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection accross parallel texts. *Natural language engineering*, 11:311–325.

R. Johansson and P. Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*.

P. Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation.

J. Lang and M. Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947, Los Angeles, California, June. Association for Computational Linguistics.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Comp. Ling.*, 19:313–330.

P. Merlo and L. van der Plas. 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet

or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore.

A. Meyers. 2007. Annotation guidelines for NomBank - noun argument structure for PropBank. Technical report, New York University.

P. Monachesi, G. Stevens, and J. Trapman. 2007. Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 77–84, Prague, Czech republic.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

S. Padó and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN*.

S. Padó. 2007. *Cross-lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

I. Titov and J. Henderson. 2007. A latent variable model for generative dependency parsing. In *Proceedings of the International Conference on Parsing Technologies (IWPT-07)*, pages 144–155, Prague, Czech Republic.

I. Titov, J. Henderson, P. Merlo, and G. Musillo. 2009. Online graph planarisation for synchronous parsing of semantic and syntactic dependencies. In *Proceedings of the twenty-first international joint conference on artificial intelligence (IJCAI-09)*, Pasadena, California, July.

L. van der Plas, T. Samardžić, and P. Merlo. 2010. Cross-lingual validity of PropBank in the manual annotation of French. In *In Proceedings of the 4th Linguistic Annotation Workshop (The LAW IV)*, Uppsala, Sweden.

D. Wu and P. Fung. 2009a. Can semantic role labeling improve SMT? In *Proceedings of the Annual Conference of European Association of Machine Translation*.

D. Wu and P. Fung. 2009b. Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of the Joint Conference of the North American Chapter of ACL/Human Language Technology*.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the International Conference on Human Language Technology (HLT)*.