Multilingual Supervision of Semantic Annotation

Peter Exner Marcus Klang Pierre Nugues

Lund University
Department of Computer Science
Lund, Sweden
{Peter.Exner, Marcus.Klang, Pierre.Nugues}@cs.lth.se

Abstract

In this paper, we investigate the annotation projection of semantic units in a practical setting. Previous approaches have focused on using parallel corpora for semantic transfer. We evaluate an alternative approach using loosely parallel corpora that does not require the corpora to be exact translations of each other. We developed a method that transfers semantic annotations from one language to another using sentences aligned by entities, and we extended it to include alignments by entity-like linguistic units. We conducted our experiments on a large scale using the English, Swedish, and French language editions of Wikipedia. Our results show that the annotation projection using entities in combination with loosely parallel corpora provides a viable approach to extending previous attempts. In addition, it allows the generation of proposition banks upon which semantic parsers can be trained.

1 Introduction

Data-driven approaches using natural language processing tackle increasingly complex tasks with ever growing scales and in more varied domains. Semantic role labeling is a type of shallow semantic parsing that is becoming an increasingly important component in information extraction (Christensen et al., 2010), question answering (Shen and Lapata, 2007), and text summarization (Khan et al., 2015).

The development of semantic resources such as FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005) made the training of models for semantic role labelers using supervised techniques possible. However, as a consequence of the considerable manual efforts needed to build proposition banks, they exist only for a few languages. An alternative approach to using supervision is to transfer knowledge between resources, a form of distant or related supervision. Methods for directly projecting semantic labels from a resource-rich language to a resource-scarce one were introduced in Padó (2007).

In this paper, we describe a method for aligning and projecting semantic annotation in loosely parallel corpora by using entities and entity-like linguistic units. Our goal is to generate multilingual PropBanks for resource-scarce languages. We used multiple language editions of Wikipedia: An English edition annotated up to a semantic level using the PropBank semantic roles, and syntactically annotated editions of Swedish and French Wikipedias. By aligning Wikipedias by entities, we constructed loosely parallel corpora and we used them to generate PropBanks in Swedish and French. We provide an evaluation of the quality of the generated PropBanks, together with an evaluation on two external FrameNets.

2 Previous Work

As an alternative to using supervised efforts for relation extraction, distant supervision can be employed to transfer relational knowledge representations from one resource to another. Distant supervision for relation extraction was introduced by Craven and Kumlien (1999) in the context of biomedical information extraction. Mintz et al. (2009) describe a method of using an external knowledge base as an indirect way of annotating text. Hoffmann et al. (2010) introduced the usage of Wikipedia infoboxes in distantly supervised relation extraction.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1007–1017, Osaka, Japan, December 11-17 2016.

The concept of transferring linguistic annotation, in the context of part-of-speech tags, across parallel corpora was introduced in Yarowsky et al. (2001). Cross-lingual annotation projection of FrameNet semantics has been described by Padó and Lapata (2009) and Basili et al. (2009). In Van der Plas et al. (2011), the authors describe an automatic method of direct transfer of PropBank semantics requiring no manual effort. Akbik et al. (2015) describe an approach to generate multilingual PropBanks using filtered annotation projection and bootstrap learning in order to handle errors stemming from translation shifts in corpora.

Most previous approaches have used professionally translated parallel corpora, mainly EuroParl (Koehn, 2005) and United Nations Corpora (Rafalovitch and Dale, 2009), to transfer semantic annotation. However, creating these resources requires manual efforts; they are thus limited in size and in the number of languages they cover. In contrast to parallel corpora, loosely parallel corpora describe similar concepts and events, but are not necessarily the result of a focused effort to translate a large corpus.

In Exner et al. (2015), we introduced the concept of using entities as a method for aligning sentences and transferring semantic content in loosely parallel corpora. However, the presented approach has the following limitations: (1) it was evaluated on one language only and (2) the evaluation was performed on the generated PropBank itself.

The contributions of this paper are the following: (1) We extend Exner et al. (2015) by including pronouns and other linguistic units that in a local context exhibit the characteristics of entities. (2) We present and evaluate two methods for aligning sentences by using entities. (3) We demonstrate the effectiveness and generalizability of our approach by projecting semantic annotations to two languages, Swedish and French, and we evaluate it using two external proposition databases, the Swedish SweFN++ (Borin et al., 2010) and French ASFALDA (Candito et al., 2014; Djemaa et al., 2016) that are both semantically-annotated corpora using adaptations of FrameNet frames. (4) We release the source code used in the annotation projection and we provide the generated PropBanks in Swedish and French[1].

3 Method

The aim of the method is to generate PropBank-like resources by fully annotating sentences in target languages using semantic content, in whole or partially, from a source language. We start with loosely parallel corpora in two languages: a **source language** (SL) expressing the semantic content that we want to transfer to a **target language** (TL). We then disambiguate and uniquely identify the entities in all the sentences. By using the unique identifier of each entity, we gain the ability to align sentences from two different languages forming sentence pairs (s_{SL}, s_{TL}). We annotate the (s_{SL}, s_{TL}) pairs, s_{SL} to semantic and syntactic levels and s_{TL} to a syntactic level. From each (s_{SL}, s_{TL}) pair, we learn the alignments between predicates (p_{SL}) in s_{SL} and verbs (v_{TL}) in s_{TL}. Finally, using the aligned entities and the predicate-verb alignments in each (s_{SL}, s_{TL}) pair, we transfer the semantic annotation in the form of predicate-argument structures. Figure 1 shows an overview of this approach.

3.1 Using Loosely Parallel Corpora

A prerequisite to projecting semantic annotation between two sentences is that they share the same semantic structure. To this end, we assumed that entities have a constraining property on the sets of predicate-argument structures they can instantiate. By aligning loosely parallel corpora through entities, pairs of sentences in two different languages that we will extract, although they are not translations of each other, should overall express the same semantic content. Furthermore, we believe that by applying our method on a large scale, the most frequent alignments of entities will elicit valid alignments.

In this context, even partial semantic content from a source sentence, s_{SL}, may be useful for annotating a target sentence, s_{TL}. As an example, consider the following sentence pair:

s_{SL} It$_{\text{A0}}$ features$_{\text{01}}$ Kelsey Grammer$_{\text{A1}}$ in his ninth ...
and is the first time$_{\text{AM-TMP}}$ the Simpsons$_{\text{A0}}$ visit$_{\text{01}}$ Italy$_{\text{A1}}$
s_{TL} I avsnittet besöker familjen Simpsons Italien
In the episode visit the family Simpsons Italy

[1] http://semantica.cs.lth.se

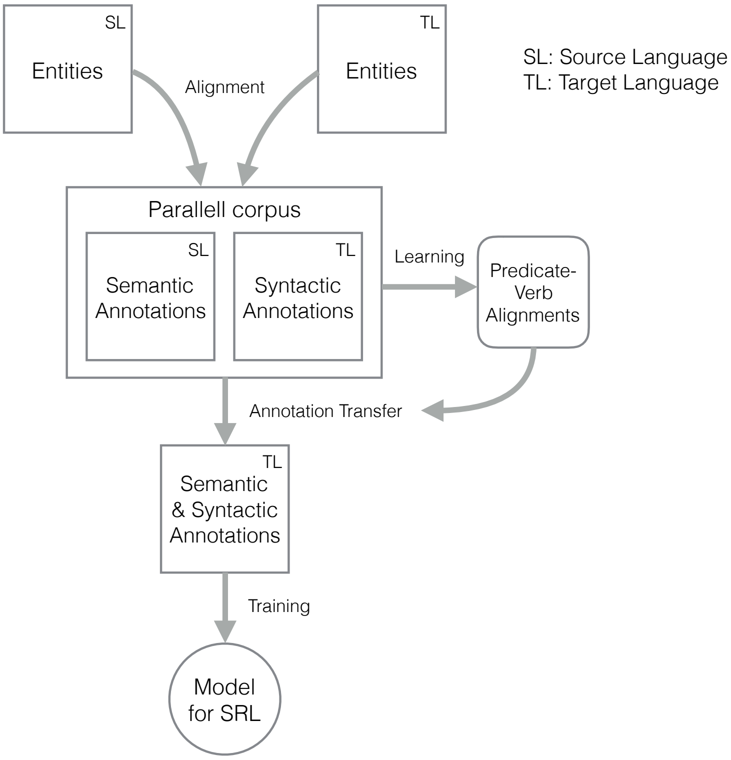

Figure 1: An overview of the approach for transferring semantic annotation from a **source language** (SL) to a **target language** (TL)

in which s_{SL} has been aligned with s_{TL} through the entities $(the\ Simpsons,\ Italy)$. s_{SL} expresses the two predicates $feature.01(it_{A0},\ Kelsey\ Grammer_{A1})$ and $visit.01(the\ first\ time_{AM-TMP},\ the\ Simpsons_{A0},\ Italy_{A1})$. Although s_{TL} is not an exact translation of s_{SL}, as it lacks the predicate $features.01$ and the temporal argument $(AM\text{-}TMP)$ of $visit.01$, the partial transfer of the semantic content enables us to annotate s_{TL} with the predicate $besöka.01(familjen\ Simpsons_{A0},\ Italien_{A1})$.

3.2 Entity Disambiguation

Entity linking is the process of finding mentions, e.g. persons, cities, organizations, events, concepts, in text, and if available, assign them with a unique identifier provided by a knowledge base. We used Wikidata Q-numbers as identifiers as they provide globally unique identifiers between the different language editions of Wikipedia. As an example, consider the following entities:

Beijing, *Pékin*, and *Pequim*

as expressed in English, French, and Portuguese respectively. Although they have differing surface forms, they are all linked to the Q956 Wikidata number, as well as in 190 other languages. In total, Wikidata covers a set of more than 13 million items that defines the entity search space.

To carry out entity linking, we reimplemented a variant of TagME (Ferragina and Scaiella, 2010). The motivating factors behind our reimplementation were:

1. It enabled us to resolve mentions to identifiers in Wikidata, providing us with multilingual and coherent entity identifiers.

2. By using the same entity linker for multiple languages, we obtained a more consistent mention resolution across all the languages.

3. It eased the adaptation to new execution environments, in our case a cluster of computing nodes.

Our implementation of TagME requires minimal grammatical information as it only needs mention statistics derived from anchors and a dictionary of mention-entity pairs and of incoming links.

The entity linking algorithm consists of four steps: detection, candidate voting, selection, and resolution of overlapping mentions.

1. We find all the possible mentions consisting of tokens in sequences up to a maximum length of 6. The mentions found at this stage might be overlapping. We treat overlapped mentions independently and they contribute votes to all the other mentions. As an example, consider the following sentence:

   Prime Minister of Japan

   containing the two mentions: *Japan* and *Prime Minister of Japan*. In this case, the overlapped mention *Japan* will contribute a vote to the overlapping mention *Prime Minister of Japan*.

2. We compute the votes for each candidate belonging to a mention. To bound the computation time, we use voting groups consisting of a collection of mentions using a sliding window approach. The vote weight per candidate is the sum of all the inlink relatedness between all the candidates (Ferragina and Scaiella, 2010). In our case, we use all the candidates in a voting group.

3. We rank all the candidates per mention using the computed votes. We then prune the mention list using a coherence criterion and a threshold that we set empirically.

4. In the final step, we resolve the mention overlap using a greedy algorithm. The algorithm selects the overlapping mention, where the entity candidate has the largest global vote, removing all the locally overlapping mentions, until there is no overlap globally.

3.3 Syntactic and Semantic Annotation

In our experimental setup, we used the English edition of Wikipedia as our SL, and we annotated it with syntactic and semantic dependencies. For the syntactic-semantic parsing, we used an open-source semantic role labeler (Choi, 2012) trained on OntoNotes 5.0 (Weischedel et al., 2013).

We transferred the semantic annotation to two TLs, the Swedish and French editions of Wikipedia, both annotated with syntactic dependencies. For French syntactic parsing, we applied a transition-based dependency parser (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012) trained on a French Treebank described in Candito et al. (2010). Correspondingly, to preprocess the Swedish edition of Wikipedia, we applied a pipeline consisting of a POS tagger (Östling, 2013) and a syntactic dependency parser (Nivre et al., 2006).

3.4 Extension to Entity-like Tokens

Entities have the property of being uniquely identifiable across languages on a global scope. However, an obvious drawback to using entities as a means of aligning sentences and transferring roles, is that roles are not always instantiated by entities. To reclaim these instances, we extended the entity alignment to include entity-like **linguistic units** (LU). We focused on units that have the property of being uniquely identifiable and limited to the scope of a sentence pair. Units correspond to sequences of tokens the entity disambiguator has either failed to classify as an entity or otherwise lack the ability to be uniquely identified in a global context.

Our algorithm detects entity-like LUs as spans of tokens sharing the same surface form in both s_{SL} and s_{TL}. In addition, we set the constraint that they occur at most once in each sentence. As a consequence, this removes any misalignment issue since a LU in s_{SL} can be matched to only one LU in s_{TL}. This method enables us to include amounts, dates, and noun phrases that the entity disambiguator fails to detect.

Using similar constraints, we also include pronouns in the detection of entity-like LUs. However, rather than using the surface form of pronouns, which would unlikely match across languages, we instead categorize them by case, gender, and number. For English, Swedish, and French, third person singular pronouns have different surface forms based on gender. Therefore, in order to increase precision, we limit the detection to only include third person pronouns. Although this constraint certainly limits the recall, this should not significantly impact the training procedure as the pronouns in the first and second persons are in very limited numbers in Wikipedia.

3.5 Aligning Sentences

The first challenge in transferring semantic annotation between loosely parallel corpora is to align sentences expressing the same semantic content. Our baseline method for aligning sentences extracts all the entities from a sentence and forms entity-sentence pairs, $(e_1...e_n, s)$. By aligning entities in different entity-sentence pairs, we form new triples containing a source sentence, a target sentence, and the subset of entities by which they are aligned $(s_{SL}, s_{TL}, e_1...e_s)$, where $k_{min} \leq s \leq k_{max}$ and k_{min}, k_{max} are prior parameters of our choice.

The baseline method is, in its simplicity, independent of any syntactic or lexical markup. It only requires the annotations from an entity disambiguator. However, one drawback lies in the inclusion of entities ungoverned by any predicate. As a consequence, the alignment of partial semantic content, as described in Sect. 3.1, becomes problematic. We therefore extended this baseline algorithm by using sets of entities projected by either arguments in s_{SL} or a verb in s_{TL}. Using this projection method, we then form entity-sentence pairs:

> $(e_1...e_p, s)$, where each entity in $(e_1...e_p)$ is governed by an argument belonging to a predicate in s_{SL}

and

> $(e_1...e_v, s)$, where each entity in $(e_1...e_v)$ is governed by a verb in s_{TL}.

The method for aligning entities in different entity-sentence pairs remains the same as for the baseline method. In Sect. 4.1, we investigate the effectiveness of the two methods under different settings.

3.6 Forming Predicate-Verb Alignments

Although we use entities as a mechanism to align sentences and transfer predicate-argument roles, predicates in s_{SL} and verbs in s_{TL} cannot be aligned by entities alone. In addition, some sentence pairs contain more than one predicate or verb, sharing the same subset of entities. This creates a combinatorial problem, where one predicate in s_{SL} could possibly be aligned to two or more verbs in s_{TL}, or vice versa. Furthermore, the application of a semantic parser to each s_{SL} annotates each predicate with a sense. This requires a method to induce new predicates and senses for the verbs in s_{TL}.

Most previous work relies on word alignments or uses bilingual dictionaries to transfer the predicate annotation between languages. However, when applied to new languages and domains, these approaches face a scaling problem requiring either training on parallel corpora or otherwise dictionaries which may not be available for every language.

Our approach builds on Exner et al. (2015) and automatically infers new predicate labels while scaling with the size of corpora and domains. A formal description of our alignment is:

1. We determine all the combinations of predicate-verb pairs, (p_i, v_k), extracted from all (s_{SL}, s_{TL}) pairs, where $p_i \in s_{SL}$ and $v_k \in s_{TL}$.

2. We assign $count(p_i, v_k)$ as the number of (p_i, v_k) in all (s_{SL}, s_{TL}), where $s_{SL} \in SL$ and $s_{TL} \in TL$.

3. For each $p_i \in SL$, we form alignments as $(p_i \rightarrow v_k) = max(count(p_i, v_1), ..., count(p_i, v_n))$.

4. For each $v_k \in (p_i \rightarrow v_k)$, we form a new TL predicate by using the lemma of v_k and an incremental counter based on the number of times v_k has appeared in an alignment.

We select the verb candidates for the alignment using lexical and syntactical rules to filter auxiliary verbs and other non-predicates.

3.7 Transferring Propositions

Given a pair of aligned sentences, (s_{SL}, s_{TL}), we transfer the semantic annotation from a predicate, $p_{SL} \in s_{SL}$, to a verb, $v_{TL} \in s_{TL}$, if $(p_{SL} \rightarrow v_{TL}) = max(count(p_i \rightarrow v_{TL})), \quad (p_i \rightarrow v_{TL}) \in (s_{SL}, s_{TL}), \quad \forall p_i \in s_{SL}$. If a s_{TL} is supervised by more than one s_{SL}, we select the s_{SL} having the larger subset of aligned entities with s_{TL}. We restrict the semantic transfer to predicate-argument structures containing at least one numbered argument and a temporal or location modifying argument, or at least two numbered arguments.

We transfer the argument roles by using the aligned entities between s_{SL} and s_{TL}. We assign the argument role to the governing token in the token span covered by each entity. However, if the argument token in s_{SL} is dominated by a preposition, we search for a preposition in s_{TL} governing the entity and assign it the argument role. We obtain the complete argument spans by taking the yield from the argument token.

4 Evaluation

In this section, we evaluate the approach described in Sect. 3 and we apply it to three language editions of Wikipedia in order to generate PropBanks for two languages: Swedish and French. The evaluation tries to answer the following questions:

1. How do different parameters and methods affect our approach?
2. What is the quality of the generated PropBanks and what level of performance can we expect in a practical setting?
3. Are there any differences between the languages, and if so what causes them?

4.1 Experimental Setup

For our experimentations, we chose the English, Swedish, and French editions of Wikipedia. These three Wikipedias are all among the top 6 in terms of article counts. As SL, we selected the English edition, and as TLs we select Swedish and French editions. We preprocessed all the articles to filter infoboxes, lists, diagrams, and to keep only text without any markup. Table 1 summarizes the statistics of the linguistic units in our chosen Wikipedias.

LANGUAGE	TOKENS	SENTENCES	ENTITIES	PREDICATES	ARGUMENTS
English	3825M	279M	439M	186M	450M
Swedish	481M	71M	58M	-	-
French	1269M	74M	181M	-	-

Table 1: Characteristics of Wikipedias used in the experimental setup

4.2 Predicate-Verb Alignment

We first evaluated how the predicate→verb alignment method described in Sect. 3.6 performs under different conditions and we examined how the number of entities, the method used, and the frequency affect the quality of the alignments. We grouped the English→Swedish alignments by their frequency into three bands: High, medium, and low. We then randomly sampled alignments from each band, in total 100 alignments and we used them to evaluate their precision. We defined precision as the number of English→Swedish alignments that we evaluate as correct divided by the total number of alignments in a sample. Figure 2 shows the precision and number of alignments using different number of entities and methods.

We observe that the precision increases with the number of entities used in the alignments. However, this increase is followed by a decrease in the number of alignments created. We also note that in all the alignments, our projection method outperforms our baseline method for aligning sentences in terms of precision. Using three projected entities, we reach a precision of roughly 80% and 1,000 alignments.

We also investigated if the higher frequency of an alignment improved precision. Figure 3 shows the breakdown of precision curves into three frequency bands, formed using projected alignments. We observe that using three projected entities, alignments with high-medium frequencies show little to no error. This provides empirical evidence to our hypothesis in Sect. 3.1, that the most frequent alignments of entities will elicit valid alignments and that precision will scale with the amount of data used by the method.

The combination of aligning sentences with three projected entities gave us the optimal trade off between precision and number of alignments created. Therefore, in the rest of the evaluation, we use these settings.

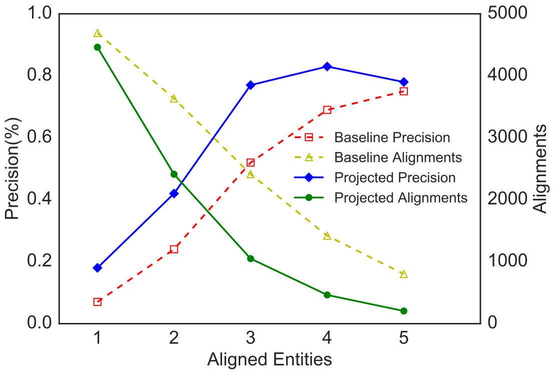

Figure 2: Graph of predicate→verb alignment precision and count under different parameter settings.

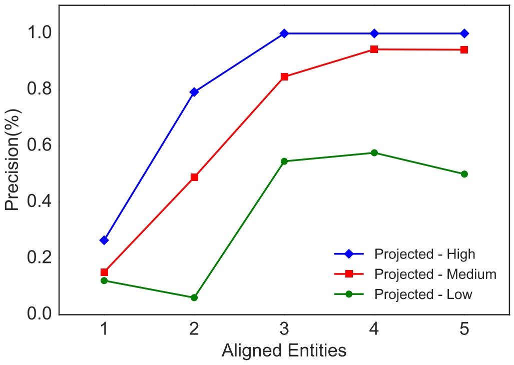

Figure 3: Graph of projected predicate→verb alignment precision, broken down by frequency band: high, medium, and low

4.3 Generated PropBanks

Using the annotation projection methods described in Sect. 3, we generated PropBanks in Swedish and French. We limited the PropBanks to only include fully annotated sentences and we removed the sentences exhibiting parsing errors, such as sentences having more than one syntactic root. We used these generated corpora to perform the error analysis in Sect. 4.5.

To evaluate our approach in a practical and automatic setting, we used samples of two linguistic resources: the Swedish FrameNet project (Borin et al., 2010) and the French FrameNet (Candito et al., 2014; Djemaa et al., 2016). We evaluated the generated Swedish and French corpora on a random sample of 100 sentences, from the Swedish FrameNet and the French FrameNet respectively. As PropBank and FrameNet have different annotation styles, we converted the sampled sentences from frame semantics to the semantics used in PropBank.

Table 2 shows the characteristics of the generated PropBanks and the FrameNets used in the evaluations.

DATASET	TOKENS	SENTENCES	PREDICATES	ARGUMENTS
Generated-Swedish	198,008	13,767	14,552	32,659
Generated-French	968,417	47,795	50,091	121,641
SweFN++ (TEST)	1,258	101	101	265
French FrameNet (TEST)	3,606	100	107	227

Table 2: Characteristics of the generated PropBanks used for training the SRL models and the FrameNets used for evaluating the trained models

4.4 Experimental Results

We evaluated the quality of the generated PropBanks in a practical setting as well as the effectiveness of using entity-like LUs in addition to entities. To assess the usefulness of the generated corpora, we first trained a semantic role labeler (Björkelund et al., 2010) on them. We split the generated corpora into 60:20:20 training, development, and testing sets, and we ran a selection process using a greedy forward selection and greedy backward elimination procedure to find the optimal set of features (Johansson and Nugues, 2008; Björkelund et al., 2009). We then used the trained models to automatically parse the test sets described in Sect. 4.3. Table 3 shows the evaluation of the semantic role labeler trained on the generated corpora.

The performance of the semantic role labeler, trained on the generated PropBanks, compares favorably with the automatic evaluations on parallel corpora described in Padó and Lapata (2009). For Swedish, using entity-like LUs, we observe an improvement of the labeled F1-measure by 10%. For French, we do not see the same dramatic increase, which we believe is caused by the large differences in pronoun classification and surface forms between English and French. We believe this discrepancy in improvement stems from projecting entity-like LUs across language groups: while English and Swedish belong to the Germanic branch, French belongs to the Romance group. Although more investigation is needed, these early results suggest that the annotation projection using entity-like LUs is most efficient when applied within a language group.

		Labeled			Unlabeled		
Language	Linguistic Units	P	R	F1	P	R	F1
Swedish	Entities (Baseline)	79.88	36.89	50.47	93.49	43.17	59.07
	Entities + Unique Tokens	84.82	44.26	58.17	92.67	48.36	63.55
	Entities + Unique Tokens + Pronouns	72.18	52.46	**60.76**	81.58	59.29	**68.67**
French	Entities (Baseline)	68.64	45.21	54.51	75.45	49.70	59.93
	Entities + Unique Tokens	64.03	48.50	55.20	70.36	53.29	**60.65**
	Entities + Unique Tokens + Pronouns	64.31	49.10	**55.69**	69.41	52.99	60.10

Table 3: Evaluation of semantic role labeling on the SweFN++ and French FrameNet corpora.

4.5 Error Analysis

To understand the quality of the generated PropBanks, we conducted an analysis of the predicate and argument errors. We randomly sampled 200 errors, of which 100 errors stemmed from the incorrect projection of argument labels and 100 were incorrect projections of predicates. Tables 4 and 5 show the type of errors for predicates and arguments respectively.

Using loosely parallel corpora, it is no surprise that the largest group of errors in predicate projection stems from sentences expressing differing semantic content. This error comes from sentence pairs, that although they contain the same subset of entities, express differing semantic content. However, as shown in Sect. 4.2, the precision of alignments increases with the number of alignments, leading us to believe that this category of error can be corrected using more data. The second largest error group is formed by different types of parsing errors occurring during the preprocessing stage. Encouragingly, only 6% of predicate projection errors stem from translation shifts, which is a further indication that entities exhibit a constraining property on the types of predicates that can instantiate them, even across languages.

Looking at argument projection errors, we again notice a group of errors stemming from misaligned sentences in loosely parallel corpora, *Differing Semantic Content* and *No Source Equivalent*. Looking beyond, alignment errors due to argument labels being assigned to the wrong token is the single most frequent error. The second largest category of errors is composed of expressions that can not be considered as entities, e.g. *In other words* and *During this time*. Finally, we observed a class of error stemming from entities undergoing a shift in specificity across sentences in two languages. These translation shifts included entities being referred to by their name in one language and by their entity type in the other language, e.g. *London→the city*.

ERROR CLASS	NUMBER
Differing Semantic Content	66
Parsing Error: Target Syntax	8
Translation Shifts: Predicate Mismatch	6
Parsing Error: Target SRL	5
Parsing Error: Entity Disambiguation	5
Auxiliary Verb	4
Light Verb Constructions	4
No Source Equivalent	1
No Target Equivalent	1
TOTAL	100

Table 4: Error analysis of English→Swedish predicate→verb alignments.

ERROR CLASS	NUMBER
Alignment Error: Non Argument Head	16
Argument is not Entity-like	14
No Source Equivalent	14
Parsing Error: SRL	14
Differing Semantic Content	13
Translation Shift: Argument Entity	12
Parsing Error: Entity Disambiguation	9
Parsing Error: Target Syntax	4
Translation Shift: Argument function	3
Parsing Error: Source Syntax	1
TOTAL	100

Table 5: Error analysis of English→Swedish argument alignments.

5 Conclusion

In this paper, we have described the construction of multilingual PropBanks by aligning loosely parallel corpora using entities. We have trained a semantic role labeler on the generated PropBanks and that we evaluated in a practical setting on frame-annotated corpora. Our results compares favorably to annotation transfer using parallel corpora. In addition, we have extended the entity alignment to include alignment by entity-like linguistic units such as pronouns and dates.

We believe the growing source of loosely parallel corpora and their alignment using entities offers an alternative way to creating multilingual hand-annotated corpora. By performing a semantic projection on loosely parallel corpora, in our case multiple language editions of Wikipedia, we have presented an alternative approach to using parallel corpora. We believe our approach can be extended beyond encyclopedias to similar resources, such as news articles in multiple languages describing the same events.

One future improvement could be to leverage ontologies that categorize entities into types. We believe that such ontologies would prove useful in adjusting the specificity of entities in order to handle some translation shifts across languages. In addition, our current method of forming predicate→verb alignments could be extended by including information about the entity type.

While projecting pronouns from English to Swedish showed an improvement, we did not observe the same improvement when projecting from English to French. Therefore, an additional avenue of investigation could compare the performance of annotation projection within versus across language groups. In addition, a coreference solver could provide an alternative means of resolving pronominal mentions to entities.

Acknowledgements

This research was supported by Vetenskapsrådet under grant 621-2010-4800, and the *Det digitaliserade samhället* and eSSENCE programs.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 397–407.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–345. Springer.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both worlds: a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465. Association for Computational Linguistics.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in swedish framenet+. In *14th EURALEX international congress*.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Seventh International Conference on Language Resources and Evaluation-LREC 2010*, pages 1840–1847. European Language Resources Association (ELRA).

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gal De Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benot Sagot, and Laure Vieu. 2014. Developing a french framenet: Methodology and first results. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Jinho D. Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. thesis, University of Colorado Boulder.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 52–60.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'99)*, pages 77–86.

Marianne Djemaa, Marie Candito, Philippe Muller, and Laure Vieu. 2016. Corpus annotation within the french framenet: a domain-by-domain methodology. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Peter Exner, Marcus Klang, and Pierre Nugues. 2015. A distant supervision approach to semantic role labeling. In *Fourth Joint Conference on Lexical and Computational Semantics (* SEM 2015)*.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Raphael Hoffmann, Congle Zhang, and Daniel S Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295.

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

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. 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: Volume 2-Volume 2*, pages 1003–1011.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Sebastian Padó. 2007. *Cross-lingual annotation projection models for role-semantic information*. Ph.D. thesis, Saarland University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Alexandre Rafalovitch and Robert Dale. 2009. United nations general assembly resolutions: A Six-Language parallel corpus. In *Proceedings of the MT Summit XII*, pages 292–299. International Association of Machine Translation, August.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 12–21, Prague, June.

Lonneke Van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.