POLYGLOT: A Crosslingual Shallow Semantic Parser

Alan Akbik
IBM Research – Almaden
650 Harry Road, San Jose, CA 95120
akbika@us.ibm.com

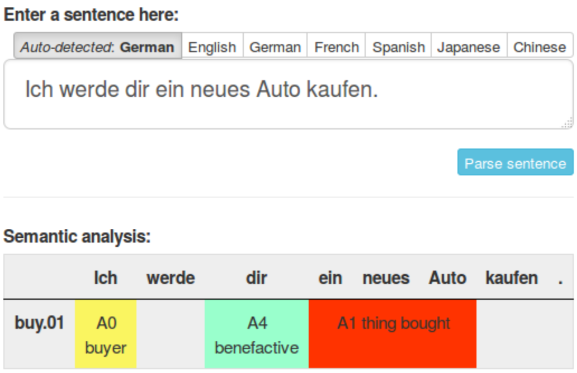
Yunyao Li
IBM Research – Almaden
650 Harry Road, San Jose, CA 95120
yunyaoli@us.ibm.com

Abstract

We present POLYGLOT, a semantic role labeling system capable of semantically parsing sentences in 9 different languages from 4 different language groups. A core differentiator is that this system predicts English Proposition Bank labels for all supported languages. This means that for instance a Japanese sentence will be tagged with the same labels as an English sentence with similar semantics would be. This is made possible by training the system with target language data that was automatically labeled with English PropBank labels using an annotation projection approach. We give an overview of our system, the automatically produced training data, and discuss possible applications and limitations of this work. We present a demonstrator that accepts sentences in English, German, French, Spanish, Japanese, Chinese, Arabic, Russian and Hindi and outputs a visualization of its shallow semantics.

1 Introduction

Semantic role labeling (SRL) is the task of labeling predicate-argument structure in sentences with shallow semantic information. One prominent labeling scheme for the English language is the Proposition Bank (Palmer et al., 2005) which annotates predicates with *frame* labels and arguments with *role* labels. Role labels roughly conform to simple questions (*Who, when, how, why, with whom*) with regards to the predicate. SRL has been found useful for downstream applications such as information extraction (IE) (Fader et al., 2011) and question answering (QA) (Shen and Lapata, 2007; Maqsud et al., 2014).



Enter a sentence here:

Ich werde dir ein neues Auto kaufen.

Semantic analysis:

	Ich	werde	dir	ein	neues	Auto	kaufen	.
buy.01	A0 buyer		A4 benefactive	A1 thing bought				

Figure 1: Example of POLYGLOT predicting English PropBank labels for a simple German sentence: The verb "*kaufen*" is correctly identified to evoke the BUY.01 frame, while "*ich*" (eng. *I*) is recognized as the *buyer*, "*ein neues Auto*" (eng. *a new car*) as the *thing bought*, and "*dir*" (eng. *for you*) as the *benefactive*.

In order to enable such parsing for languages other than English, annotation efforts are under way to create Proposition Bank-style resources for some other languages such as Chinese (Xue and Palmer, 2005) and Hindi (Bhatt et al., 2009). However, such efforts are generally costly in terms of time and experts required, making it difficult to expand SRL to new target languages.

In our research, we investigate a different approach based on *annotation projection* (Padó and Lapata, 2009; Van der Plas et al., 2011) to automatically generate such resources for arbitrary target languages. The idea is to utilize collections of parallel text to transfer predicted SRL labels from English sentences onto target language translations. We presented a method of filtered projection and sampled learning in order to limit errors stemming from non-literal translations (Akbik et al., 2015). With our approach, we auto-generated Proposition Bank-style resources for 7 languages, namely Arabic, Chinese, French, German, Hindi, Russian and Spanish.

A core difference between these generated PropBanks and manual annotation efforts is that we re-use English Proposition Bank labels for the target languages. This enables us to train an SRL system to consume text in various languages and make predictions in a shared semantic label set. Refer to Figure 1 for an illustration of a German-language sentence tagged with English PropBank labels by our system.

Following such an approach, we believe that English PropBank labels might eventually become a basis of "universal" shallow semantic labels similar to the way that Stanford dependencies are the basis of universal dependencies (De Marneffe et al., 2014). However, one key question is to what degree English PropBank frame and role labels are appropriate for target language shallow semantics and in how far such an approach can handle language-specific phenomena or semantic concepts.

Contributions. In order to facilitate discussion of such questions with the research community, we present POLYGLOT, an SRL system trained on auto-generated PropBanks. Given a sentence in one of 9 languages, the system visualizes a shallow semantic parse with predicted English PropBank labels. The demonstrator allows us to illustrate our envisioned approach of parsing different languages into a shared shallow semantic abstraction. It also enables researchers to experiment with the tool to understand the breadth of shallow semantic concepts currently covered, and to discuss limitations and the potential of such an approach for downstream applications such as multilingual information extraction and question answering.

The rest of this paper is structured as follows: We briefly revisit work in annotation projection to illustrate how the training data was created and what data sets were used to train the system. We then give a tour of the demonstrator and discuss directions for future research.

2 Annotation Projection

Annotation projection takes as input a word-aligned parallel corpus of English sentences and their translations in a target language. A syntactic parser and a semantic role labeler then produce labels for the English sentences. In a projection step, these labels are transferred along word alignments onto the target language side. The underlying theory is that translated sentence pairs share a degree of semantic similarity, making such projection possible (Padó and Lapata, 2009).

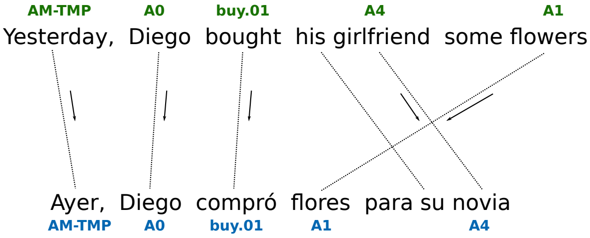

AM-TMP
A0
buy.01
A4
A1
Yesterday, Diego bought his girlfriend some flowers
Ayer, Diego compró flores para su novia
AM-TMP
A0
buy.01
A1
A4

Figure 2: A word aligned English-Spanish sentence pair. SRL is used to predict labels for the English side (green labels). The labels are then projected onto aligned words in the target language side, thereby automatically labeling the Spanish sentence with English PropBank labels (blue labels).

Refer to Figure 2 for an example with a pair of very simple sentences: The English sentence is labeled with the appropriate frame (BUY.01) and role labels: "*Diego*" is labeled as the *buyer* (**A0** in PropBank annotation), "*some flowers*" as the *thing bought* (**A1**) and "*his girlfriend*" as the *benefactive* (**A4**). In addition, "yesterday" is labeled **AM-MOD**, signifying a temporal context of this frame. These labels are then projected onto aligned Spanish words. For instance, "*compro*" is word-aligned to "*bought*" and is therefore labeled as BUY.01. This produces a Spanish sentence labeled with English PropBank labels that can in turn be used to train an SRL system for Spanish.

State-of-the-art. Previous work analyzed errors in annotation projection and found that they are often caused by non-literal translations (Akbik et al., 2015). For this reason, previous work defined lexical and syntactic constraints to increase projection quality; These include verb filters to allow only verbs to be labeled as frames (Van der Plas et al., 2011), heuristics that ensure that only heads of syntactic constituents are labeled as arguments (Padó and Lapata, 2009) and the use of verb translation dictionaries to guide frame mappings. In (Akbik et al., 2015), we additionally proposed a process of filtered projection and sampled learning, and executed the approach to create Proposition Banks for 7 target languages.

We evaluated the proposed approach and found the quality of the generated PropBanks to be moderate to high, depending on the target language. Table 1 shows estimated precision, recall and F1-score for each language respectively with two evaluation methods. Partial evaluation counts correctly labeled incomplete constituents as true positives while exact evaluation only counts correctly labeled complete constituents as true positives. Please refer to (Akbik et al., 2015) for a full breakdown of evaluation results and error sources.

In order to generate training data for POLY-

		PREDICATE			ARGUMENT				
LANG.	Match	P	R	F1	P	R	F1	Agr	κ
Arabic	part.	0.97	0.89	0.93	0.86	0.69	0.77	0.92	0.87
	exact	0.97	0.89	0.93	0.67	0.63	0.65	0.85	0.77
Chinese	part.	0.97	0.88	0.92	0.93	0.83	0.88	0.95	0.91
	exact	0.97	0.88	0.92	0.83	0.81	0.82	0.92	0.86
French	part.	0.95	0.92	0.94	0.92	0.76	0.83	0.97	0.95
	exact	0.95	0.92	0.94	0.86	0.74	0.8	0.95	0.91
German	part.	0.96	0.92	0.94	0.95	0.73	0.83	0.95	0.91
	exact	0.96	0.92	0.94	0.91	0.73	0.81	0.92	0.86
Hindi	part.	0.91	0.68	0.78	0.93	0.66	0.77	0.94	0.88
	exact	0.91	0.68	0.78	0.58	0.54	0.56	0.81	0.69
Russian	part.	0.96	0.94	0.95	0.91	0.68	0.78	0.97	0.94
	exact	0.96	0.94	0.95	0.79	0.65	0.72	0.93	0.89
Spanish	part.	0.96	0.93	0.95	0.85	0.74	0.79	0.91	0.85
	exact	0.96	0.93	0.95	0.75	0.72	0.74	0.85	0.77

Table 1: Estimated precision and recall over seven languages from an earlier evalution.

LANG.	DEP. PARSER	DATA SETS	#SENT.
Arabic	STANFORD	`UN`, `OpenSubs2016`	24,5M
Chinese	MATE-G	`UN`, `OpenSubs2016`	12,2M
French	MATE-T	`UN`, `OpenSubs2016`	36M
German	MATE-T	`Europarl`, `OpenSubs2016`	14,1M
Hindi	MALT	`Hindencorp`	54K
Japanese	JJST	`Tatoeba`, `OpenSubs2016`	1,7M
Russian	MALT	`UN`, `OpenSubs2016`	22,7M
Spanish	MATE-G	`UN`, `OpenSubs2016`	52,4M

Table 2: Parsers and datasets used to generate training data for each language.

Dependency parsers: STANFORD: (Green and Manning, 2010), MATE-G: (Bohnet, 2010), MATE-T: (Bohnet and Nivre, 2012), MALT: (Nivre et al., 2006). **Parallel corpora**: `Tatoeba`: (Tiedemann, 2012), `OpenSubs2016`: (Lison and Tiedemann,), `UN`: (Rafalovitch et al., 2009), `Europarl`: (Koehn, 2005), `Hindencorp`: (Bojar et al., 2014). **Word alignment**: The `UN` corpus is already word-aligned. For others, we use the Berkeley Aligner (DeNero and Liang, 2007).

GLOT, we used the approach described in (Akbik et al., 2015) with additional sources of parallel data. One very large additional source of parallel data is the `OpenSubs2016`corpus automatically mined from movie subtitles for a large range of languages (Lison and Tiedemann,). We also added the Japanese language as an additional target language, with the `Tatoeba`parallel data set extracted from language learning examples[1]. Both data sets were obtained from the OPUS project (Tiedemann, 2012). Table 2 gives an overview of all parallel datasets and dependency parsers we used.

3 POLYGLOT Demonstrator

The demonstrator is a Web interface for SRL over the 8 target languages listed in Table 2 as well as English. Users begin by entering a sentence in the text field and clicking the "parse sentence" button. The system employs a language-detection module so that users do not need to explicitly select their input language. However, in rare cases where language-detection fails, users can also select the input language by toggling the appropriate button in the upper right corner.

A language-specific NLP pipeline is first used to preprocess a submitted sentence. This includes tokenization, lemmatization, morphological analysis, part-of-speech tagging and dependency parsing (see Table 2). The preprocessed sentence is then passed to an instantiation of our SRL system trained using the generated PropBanks for each target language, and trained using the CONLL09 shared dataset for English.

Output. The syntactic and semantic parsing results are displayed below the input field. The topmost result row is the semantic analysis, presented as a grid in which each row corresponds to one identified semantic frame. The grid highlights sentence constituents labeled with roles and includes role descriptions for better interpretability of the parsing results.

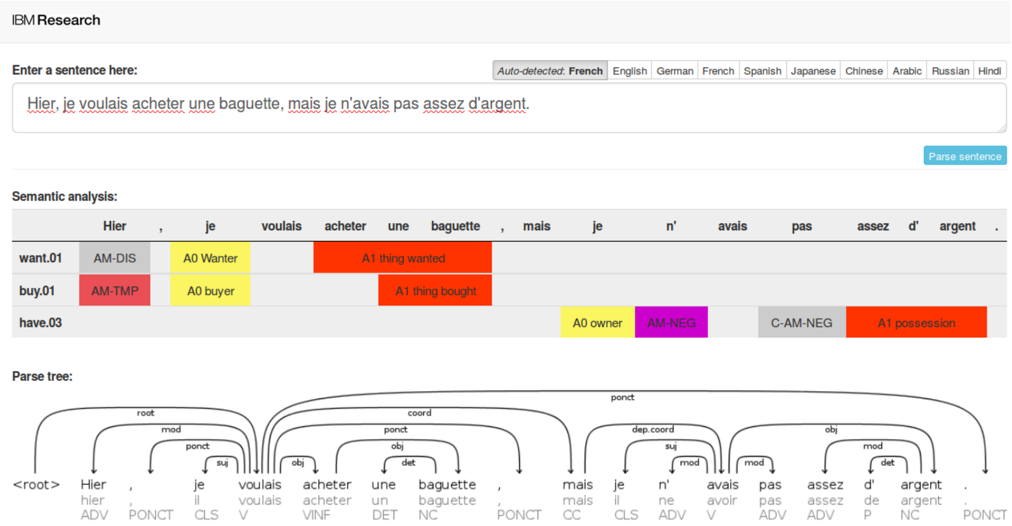
Below the results of the semantic analysis are two rows for more detailed inspection of the parsing results. The first is a visualization of the dependency parse tree generated using WHATSWRONGWITHMYNLP[2], while the second (not shown in Figure **??**) is the full syntactic-semantic parse in CONLL format, including morphological information and other features not present in the dependency tree visualizaion. This output is helpful to identify SRL errors that are caused by mistakes made earlier in the NLP pipeline. A common example are errors in dependency parsing that cause incorrect constituents to be selected as arguments.

Example sentence. In Figure **??**, we illustrate the result visualization of the tool. The user has entered the French sentence "*Hier, je voulais acheter une baguette, mais je n'avais pas assez d'argent*" (engl. "Yesterday I wanted to buy a baguette but I didn't have enough money") into the text field. As indicated in the top left corner, French is auto-detected as the language of interest.

The semantic analysis is displayed below the input field. The first column in the grid indicates that three frames have been identified: WANT.01,

[1] http://tatoeba.org/eng/

[2] https://code.google.com/archive/p/whatswrong/



IBM Research

Enter a sentence here:

Auto-detected: **French** | English | German | French | Spanish | Japanese | Chinese | Arabic | Russian | Hindi

Hier, je voulais acheter une baguette, mais je n'avais pas assez d'argent.

Parse sentence

Semantic analysis:

	Hier	,	je	voulais	acheter	une	baguette	,	mais	je	n'	avais	pas	assez	d'	argent	.
want.01	AM-DIS		A0 Wanter		A1 thing wanted												
buy.01	AM-TMP		A0 buyer			A1 thing bought											
have.03										A0 owner	AM-NEG		C-AM-NEG	A1 possession			

Parse tree:

Figure 3: POLYGLOT's Web UI with a French example sentence.

BUY.01 and HAVE.03. The second row in the grid corresponds to the WANT.01 frame, which identifies "*je*" (eng. "I") as the *wanter* and "*acheter un baguette*" (eng. "*buy a baguette*") as the *thing wanted*. The arguments are color-coded by PropBank argument type for better readability. For instance, in the PropBank annotation scheme, the agents of the three frames in the example (the *wanter*, the *buyer* and the *owner*) are all annotated with the same role (**A0**). They are thus highlighted in the same yellow color in the visualization. This allows a user to quickly gauge whether the semantic analysis of the sentence is correct.

4 Discussion

We are currently pursuing several lines of work to investigate the questions raised in the introduction to this paper.

Evaluate POLYGLOT in downstream application. A potential advantage of this is that downstream applications such as IE or QA could be designed solely against this abstraction. For instance, one might seek to extract a list of planned acquisitions from a multi-language corpus. Similar to the design of previous QA systems (Maqsud et al., 2014), this query could be expressed in shallow semantics. Since we use the same tagset across all languages, the query can match constituents marked as *thing bought* in arbitrary languages, making our information extraction system multilingual without additional effort.

Increasing the quality of the training data. Our current work on annotation projection focuses on further increasing the quality of the generated data. We are investigating additional heuristics for filtering and auto-tagging target language data; for instance, certain role labels in the English PropBank are more syntactic than semantic in nature, such as the AM-MOD label for auxiliary verbs and AM-NEG for negations. Our current work indicates that it is better to use simple syntactic rules for target language sentenes to annotate such constituents than trying to project such information. We are also experimenting with heuristics to address language-specific problems. For instance, seperable prefixes change the meanings of German verbs and must therefore be explicitly included in verb lemmatization even though they can be separated into distinct tokens[3].

Investigate impact of universal dependencies.

5 Conclusion

References

[Akbik et al.2015] Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *ACL 2015, 53rd Annual Meeting of the Association for Computational Linguistics Beijing, China*, page to appear.

[Bhatt et al.2009] Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

[3] The verb "vorhaben" for instance contains the separable prefix "vor" which in V2 order will be placed as a separate token at the end of a phrase.

[Bohnet and Nivre2012] 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.

[Bohnet2010] Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

[Bojar et al.2014] Ondřej Bojar, Vojtěch Diatka, Pavel Rychlỳ, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, Daniel Zeman, et al. 2014. Hindencorp–hindi-english and hindi-only corpus for machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

[De Marneffe et al.2014] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

[DeNero and Liang2007] John DeNero and Percy Liang. 2007. The berkeley aligner. `http://code.google.com/p/berkeleyaligner/`.

[Fader et al.2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

[Green and Manning2010] Spence Green and Christopher D Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402. Association for Computational Linguistics.

[Koehn2005] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

[Lison and Tiedemann] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

[Maqsud et al.2014] Umar Maqsud, Sebastian Arnold, Michael Hülfenhaus, and Alan Akbik. 2014. Nerdle: Topic-specific question answering using wikia seeds. In *COLING (Demos)*, pages 81–85.

[Nivre et al.2006] 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.

[Padó and Lapata2009] Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

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

[Rafalovitch et al.2009] Alexandre Rafalovitch, Robert Dale, et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

[Shen and Lapata2007] Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21.

[Tiedemann2012] Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation, MAY 21-27, 2012, Istanbul, Turkey*, pages 2214–2218.

[Van der Plas et al.2011] 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.

[Xue and Palmer2005] Nianwen Xue and Martha Palmer. 2005. Automatic semantic role labeling for chinese verbs. In *IJCAI*, volume 5, pages 1160–1165. Citeseer.