POLYGLOT: Multilingual Semantic Role Labeling with Unified Labels

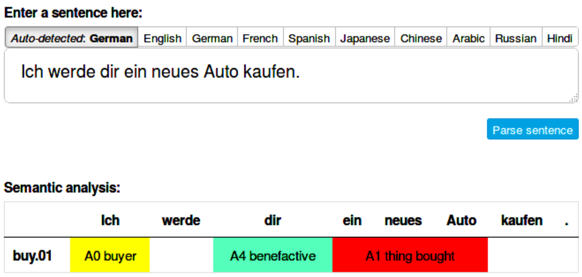
Alan Akbik **Yunyao Li**
IBM Research
Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA
{akbika,yunyaoli}@us.ibm.com

Abstract

Semantic role labeling (SRL) identifies the predicate-argument structure in text with semantic labels. It plays a key role in understanding natural language. In this paper, we present POLYGLOT, a multilingual semantic role labeling system capable of semantically parsing sentences in 9 different languages from 4 different language groups. The core of POLYGLOT are SRL models for individual languages trained with automatically generated Proposition Banks (Akbik et al., 2015). The key feature of the system is that it treats the semantic labels of the English Proposition Bank as "universal semantic labels": Given a sentence in any of the supported languages, POLYGLOT applies the corresponding SRL and predicts English PropBank frame and role annotation. The results are then visualized to facilitate the understanding of multilingual SRL with this unified semantic representation.

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, what, when, where, how much, with whom*) with regards to the predicate. SRL is important for understanding natural language; it has been found useful for many applications such as information extraction (Fader et al., 2011) and question answering (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*" (*I*) is recognized as the *buyer*, "*ein neues Auto*" (*a new car*) as the *thing bought*, and "*dir*" (*for you*) as the *benefactive*.

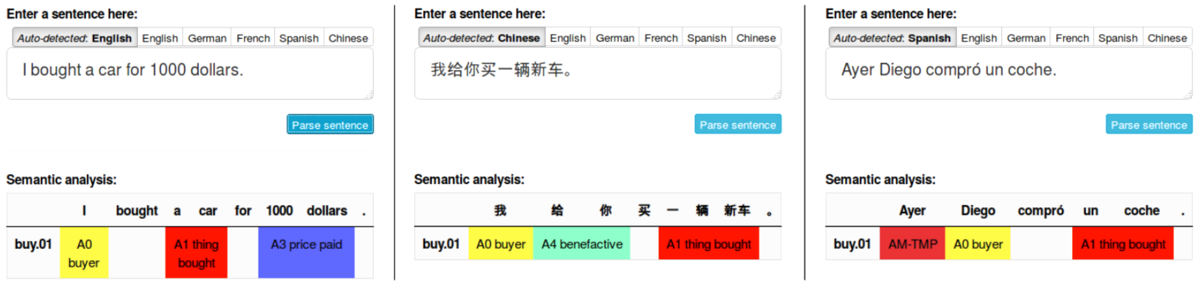
Not surprisingly, enabling SRL for languages other than English has received increasing attention. One relevant key effort is to create Proposition Bank-style resources for different languages, such as Chinese (Xue and Palmer, 2005) and Hindi (Bhatt et al., 2009). However, the conventional approach of manually generating such resources is costly in terms of time and experts required, hindering the expansion of SRL to new target languages.

An alternative approach is *annotation projection* (Padó and Lapata, 2009; Van der Plas et al., 2011) that utilizes parallel corpora to transfer predicted SRL labels from English sentences onto sentences in a target language. It has shown great promise in automatically generating such resources for arbitrary target languages. In previous work, we presented an approach based on filtered projection and bootstrapped learning to auto-generate Proposition Bank-style resources for 7 languages, namely Arabic, Chinese, French, German, Hindi, Russian and Spanish (Akbik et al., 2015).

Unified semantic labels across all languages.

One key difference between auto-generated PropBanks and manually created ones is that the former use English Proposition Bank labels for

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 1–6,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics



Enter a sentence here:

I bought a car for 1000 dollars.

Semantic analysis:

	I	bought	a	car	for	1000	dollars	.
buy.01	A0 buyer		A1 thing bought			A3 price paid		

Enter a sentence here:

我给你买一辆新车。

Semantic analysis:

	我	给	你	买	一	辆	新车	。
buy.01	A0 buyer	A4 benefactive			A1 thing bought			

Enter a sentence here:

Ayer Diego compró un coche.

Semantic analysis:

	Ayer	Diego	compró	un	coche	.
buy.01	AM-TMP	A0 buyer		A1 thing bought		

Figure 2: Side by side view of English, Chinese and Spanish sentences parsed in POLYGLOT's Web UI. English PropBank frame and role labels are predicted for all languages: All example sentences evoke the BUY.01 frame and have constituents accordingly labeled with roles such as the *buyer*, the *thing bought*, the *price paid* and the *benefactive*.

all target languages, while the latter use language-specific labels. As such, the auto-generated PropBanks allow us to train an SRL system to consume text in various languages and make predictions in a shared semantic label set, namely English PropBank labels. Refer to Figures 1 and 2 for examples of how our system predicts frame and role labels from the English Proposition Bank for sentences in German, English, Chinese and Spanish.

Similar to how Stanford dependencies are one basis of *universal dependencies* (De Marneffe et al., 2014; Nivre, 2015), we believe that English PropBank labels have the potential to eventually become a basis of "universal" shallow semantic labels. Such a unified representation of shallow semantics, we argue, may facilitate applications such as multilingual information extraction and question answering, much in the same way that universal dependencies facilitate tasks such as crosslingual learning and the development and evaluation of multilingual syntactic parsers (Nivre, 2015). The key questions, however, are (1) to what degree English PropBank frame and role labels are appropriate for different target languages; and (2) how far this approach can handle language-specific phenomena or semantic concepts.

Contributions. To facilitate the discussions of the above questions, we present POLYGLOT, an SRL system trained on auto-generated PropBanks for 8 languages plus English, namely Arabic, Chinese, French, German, Hindi, Japanese, Russian and Spanish. Given a sentence in one of these 9 languages, the system applies the corresponding SRL and visualizes the shallow semantic parse with predicted English PropBank labels. POLYGLOT allows us to illustrate our envisioned approach of parsing different languages into a shared shallow semantic abstraction based on the English Proposition Bank. 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.

2 System Overview

Figure 3 depicts the overall architecture of POLYGLOT. First, we automatically generate labeled training data for each target language with annotation projection (Akbik et al., 2015) (Figure 3 Step 1). We use the labeled data to train for each language an SRL system (Figure 3 Step 2) that predicts English PropBank frame and role labels. Both the creation of the training data and the training of the SRL instances are one-time processes.

POLYGLOT provides a Web-based GUI to allow users to interact with the SRL systems (Figure 3 Step 3). Given a natural language sentence, depending on the language of the input sentence, POLYGLOT selects the appropriate SRL instance and displays on the GUI the semantic parse, as well as syntactic information.

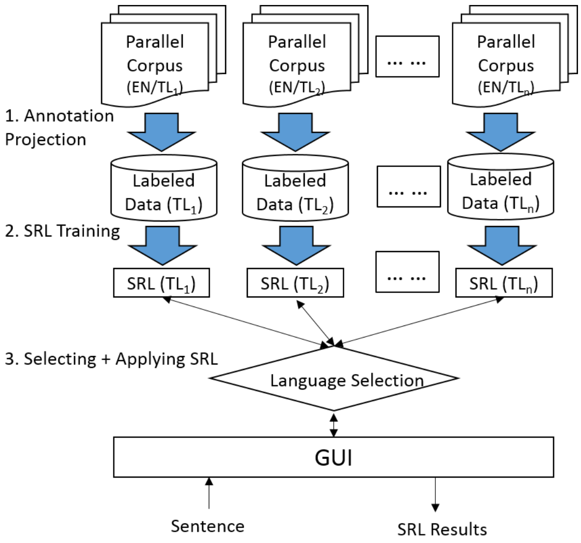


Figure 3: System overview

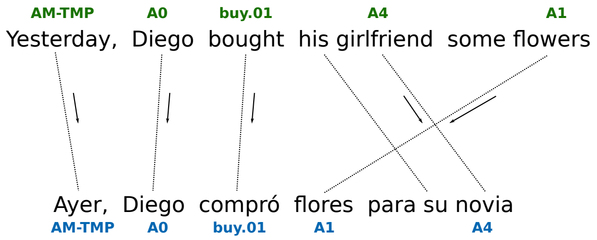

AM-TMP
A0
buy.01
A4
A1
Yesterday, Diego bought his girlfriend some flowers
Ayer, Diego compró flores para su novia
AM-TMP
A0
buy.01
A1
A4

Figure 4: Annotation projection for a word-aligned English-Spanish sentence pair.

In the next sections, we briefly describe the creation of the labeled data and the training of the SRL systems, followed by a tour of the Web UI.

3 Auto-Generation of Labeled Data

We followed an annotation projection approach to automatically generate the labeled data for different languages. This approach takes as input a word-aligned parallel corpus of English sentences and their translations in a target language (TL). A semantic role labeler then predicts labels for the English sentences. In a projection step, these labels are transferred along word alignments onto the target language sentences. The underlying theory is that translated sentence pairs share a degree of semantic similarity, making such projection possible (Padó and Lapata, 2009).

Figure 4 illustrates an example of annotation projection: Using an SRL system trained with the English Proposition Bank, the English sentence is labeled with the appropriate frame (BUY.01) and role labels: "*Diego*" 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-TMP**, signifying a temporal context of this frame. These labels are then projected onto the aligned Spanish words. For instance, "*compró*" is word-aligned to "*bought*" and thus labeled as BUY.01. The projection produces a Spanish sentence labeled with English PropBank labels; such data can in turn be used to train an SRL system for Spanish.

State-of-the-art. Direct annotation projection often introduces errors, mostly due to non-literal translations (Akbik et al., 2015). Previous work defined lexical and syntactic constraints to increase projection quality, such as 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 bootstrapped learning, and successfully created Proposition Banks for 7 target languages. We 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 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. For more details we refer the reader to (Akbik et al., 2015) .

4 Semantic Role Labeling

Using the auto-generated labeled data, we train the semantic role labeler of the MATE toolkit (Björkelund et al., 2009), which achieved state-of-the-art semantic F1-score in the multilingual semantic role labeling task of the CoNLL-2009 shared task (Hajič et al., 2009). The parser is implemented as a sequence of local logistic regression classifiers for the four steps of predicate identification, predicate classification, argument identification and argument classification. In addition, it implements a global reranker to rerank sets of local predictions. It uses a standard feature set of lexical and syntactic features.

Preprocessing. Before SRL, we execute a pipeline of NLP tools to extract the required lexical, morphological and syntactic features. To facilitate reproducability of the presented work, we use publicly available open source tools and pre-

		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 our previous evaluation (Akbik et al., 2015).

Language	NLP Preprocessing	Parallel Data Sets	#Sentences
Arabic	StanfordCoreNLP, KhojaStemmer, StanfordParser	UN, OpenSubtitles	24,5M
Chinese	StanfordCoreNLP, MateParser	UN, OpenSubtitles	12,2M
English	ClearNLP	n/a	n/a
French	StanfordCoreNLP, MateTransitionParser	UN, OpenSubtitles	36M
German	StanfordCoreNLP, MateTransitionParser	Europarl, OpenSubtitles	14,1M
Hindi	TnTTagger, MaltParser	Hindencorp	54K
Japanese	JJST	Tatoeba, OpenSubtitles	1,7M
Russian	TreeTagger, MaltParser	UN, OpenSubtitles	22,7M
Spanish	StanfordCoreNLP, MateParser	UN, OpenSubtitles	52,4M

Table 2: NLP tools and source of parallel data used for each language. Since English is the source language for annotation projection, no parallel data was required to train SRL.
NLP tools: StanfordCoreNLP: (Manning et al., 2014) , TnTTagger: (Brants, 2000), TreeTagger: (Schmid, 1994), KhojaStemmer: (Khoja and Garside, 1999), StanfordParser: (Green and Manning, 2010), StanfordCoreNLP: (Choi and McCallum, 2013), MateParser: (Bohnet, 2010), JJST: *proprietary system*, MateTransitionParser: (Bohnet and Nivre, 2012), MaltParser: (Nivre et al., 2006).

trained models where available. A breakdown of the preprocessing tools used for each language is given in Table 2.

Data sets. In order to generate training data for Polyglot, we used the following sources of parallel data: The `UN` corpus of official United Nations documents (Rafalovitch et al., 2009), the `Europarl` corpus of European parliament proceedings (Koehn, 2005), the `OpenSubtitles` corpus of movie subtitles (Lison and Tiedemann, 2016), the `Hindencorp` corpus automatically gathered from web sources (Bojar et al., 2014) and the `Tatoeba` corpus of language learning examples[1]. The data sets were obtained from the OPUS project (Tiedemann, 2012) and word aligned using the Berkeley Aligner[2]. Table 2 lists the data sets used for each language and the combined number of available parallel sentences.

5 Polyglot User Interface

The Web-based GUI of Polyglot allows users to enter sentences in one of 9 languages and request a shallow semantic analysis. Figure 5 presents a screenshot of the GUI. Users begin by entering a sentence in the text field and clicking on the "parse sentence" button. As indicated in Figure 3, it is then passed to a language-specific NLP pipeline based on the associated language that is detected automatically by default or specified by the user. The pipeline tokenizes and lemmatizes the sentence, performs morphological analysis, dependency parsing and semantic role labeling.

Output. The syntactic and semantic parsing results are displayed below the input field, following the design of (Björkelund et al., 2010): The topmost result table 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 the GUI shows two more detailed views of the parsing results. The first visualizes the dependency parse tree generated using WhatsWrongWithMyNLP[3], while the second (omitted in Figure 5 in the interest of space) displays the full syntactic-semantic parse in CoNLL format, including morphological information and other features not present in the dependency tree visualization. These two views may be helpful to users that wish to identify possible sources of SRL errors. For instance, a common error class stem from errors in dependency parsing, causing incorrect constituents to be labeled as arguments.

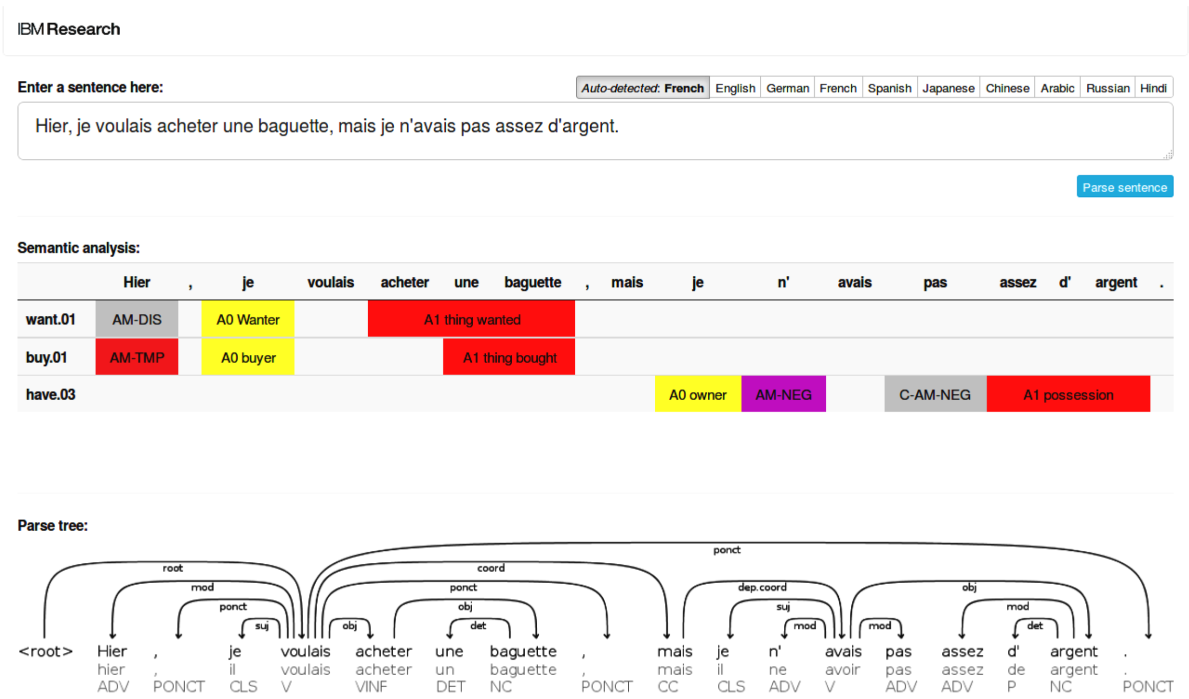
Example sentence. Figure 5 illustrates the result visualization of the tool. A user enters a sentence "*Hier, je voulais acheter une baguette, mais je n'avais pas assez d'argent*" (engl. "*Yesterday I wanted to buy a baguette but I didn't have enough money*") into the text field. As indicated in the top left corner, this sentence is auto-detected to be French.

The results of semantic analysis is displayed below the input field. The first column in the grid indicates that three frames have been identified: WANT.01, BUY.01 and HAVE.03. The second row in the grid corresponds to the WANT.01 frame, which identifies "*je*" (engl. "*I*") as the *wanter* and "*acheter une baguette*" (engl. "*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,

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

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

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



IBM Research

Enter a sentence here:

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

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:

<root>	Hier	,	je	voulais	acheter	une	baguette	,	mais	je	n'	avais	pas	assez	d'	argent	.
	hier	,	il	voulais	acheter	un	baguette	,	mais	il	ne	avoir	pas	assez	de	argent	.
	ADV	PONCT	CLS	V	VINF	DET	NC	PONCT	CC	CLS	ADV	V	ADV	ADV	P	NC	PONCT

Figure 5: POLYGLOT's Web UI with a French example sentence.

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].

6 Demonstration and Outlook

We present POLYGLOT as a hands-on demo where users can enter sentences and request shallow semantic analyses. We also plan to make it publicly accessible in the future. We are currently working on improving the annotation projection approach to generate higher quality data by experimenting with further constraints, language-specific heuristics and improved frame mappings. We are particularly interested in how far English Proposition Bank labels are suitable for arbitrary target languages and may serve as basis of a "universal semantic role labeling" framework: we are qualitatively analysing auto-generated PropBanks in comparison to manual efforts; meanwhile, we are evaluating POLYGLOT in downstream applications such as multilingual IE. Through the presentation of POLYGLOT, we hope to engage the research community in this discussion.

[4] The example sentence in Figure 5 contains one error: The word "*hier*" (engl. "*yesterday*") should be labeled **AM-TMP** instead of **AM-DIS**. All other labels are correct.

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. In *ACL 2015, 53rd Annual Meeting of the Association for Computational Linguistics Beijing, China*, page to appear.

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.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth CoNLL: Shared Task*, pages 43–48, Boulder, Colorado, June. 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 Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 EMNLP-CoNLL*, pages 1455–1465. Association for Computational Linguistics.

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

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 LREC*.

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics.

Jinho D. Choi and Andrew McCallum. 2013. Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

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.

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.

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

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth CoNLL: Shared Task*, pages 1–18. Association for Computational Linguistics.

Shereen Khoja and Roger Garside. 1999. Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

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

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

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.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

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

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, 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.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of LREC, MAY 21-27, 2012, Istanbul, Turkey*, pages 2214–2218.

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.

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