Any-language frame-semantic parsing

Anders Johannsen, Héctor Martínez Alonso, Anders Søgaard
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
{ajohannsen,alonso,soegaard}@hum.ku.dk

Abstract

We present a multilingual corpus of Wikipedia and Twitter texts annotated with FRAMENET 1.5 semantic frames in nine different languages, as well as a novel technique for weakly supervised cross-lingual frame-semantic parsing. Our approach only assumes the existence of linked, comparable source and target language corpora (e.g., Wikipedia) and a bilingual dictionary (e.g., Wiktionary or BABELNET). Our approach uses a truly interlingual representation, enabling us to use the *same* model across all nine languages. We present average error reductions over running a state-of-the-art parser on word-to-word translations of 46% for target identification, 37% for frame identification, and 14% for argument identification.

1 Introduction

Frame-semantic parsing is the task of automatically finding semantically salient targets in text, disambiguating the targets by assigning a sense (frame) to them, identifying their arguments, and labeling these arguments with appropriate roles. The FRAMENET 1.5 lexicon[1] provides a fixed repository of semantic frames and roles, which we use in the experiments below.

Several learning and parsing algorithms have been developed for frame-semantic analysis (Johansson and Nugues, 2007; Das et al., 2014; Täckström et al., 2015), and frame semantics has been successfully applied to question-answering (Shen and Lapata, 2007), information extraction (Surdeanu et al., 2003) and knowledge extraction (Søgaard et al., 2015b).

[1] https://framenet.icsi.berkeley.edu/

In contrast to Propbank-style semantic-role labeling (Titov and Klementiev, 2012), only very limited frame-semantic resources exist for languages other than English. We therefore focus on *multilingual* or cross-language frame-semantic parsing, leveraging resources for English and other major languages to build *any*-language parsers. We stress that we learn frame-semantic parsing models that can be applied to *any* language, rather than cross-lingual transfer models for specific target languages. Our approach relies on inter-lingual word embeddings (Søgaard et al., 2015a), which are built from topic-aligned documents. Word embeddings have previously been used for monolingual frame-semantic parsing by Hermann et al. (2014).

Contributions This paper makes the following three contributions. We present a new multilingual frame-annotated corpus covering five topics, two domains (Wikipedia and Twitter), and nine languages. We implement a simplified version of the frame-semantic parser introduced in Das et al. (2014). Finally, we show how to modify this parser to learn any-language frame-semantic parsing models using inter-lingual word embeddings (Søgaard et al., 2015a).

2 Data annotation

Figure 1 depicts a FRAMENET 1.5 frame-semantic analysis of a German sentence from Wikipedia. The annotator marked two words, *Idee* and *kam*, as targets. In frame-semantic parsing, target identification is the task of deciding which words (i.e. targets) trigger FRAMENET frames. Frame identification is the problem of disambiguating targets by labeling them with frames, e.g., COGITATION or COMING_UP_WITH. Argument identification is the problem of identifying the arguments of frames, e.g., *Idee* for COMING_UP_WITH.

We had linguistically trained students anno-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2062–2066,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

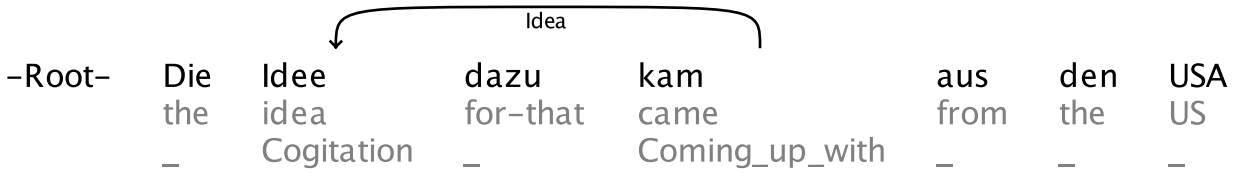

Figure 1: Frame semantic annotation from the German Wikipedia data (Women's Rights)

tate about 200 sentences from Wikipedia and 200 tweets each in their native language. The data was pre-annotated by obtaining all English translation equivalents of the source language words through BABELNET[2], finding associated frames in the FRAMENET 1.5 training data. We presented annotators with *all* frames that could be triggered by any of the target word's translations. Both data from Wikipedia and Twitter cover the same five topics: Google, Angelina Jolie, Harry Potter, Women's Rights, and Christiano Ronaldo. The topics were chosen to guarantee coverage for all nine languages, both in Wikipedia and Twitter. Our corpus, which covers nine languages, is publicly available at `https://github.com/andersjo/any-language-frames` The languages we cover are Bulgarian (BG), Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), French (FR), Italian (IT) and Swedish (SV). English is included as a sanity check of our cross-lingual annotation setup.

The English, Danish, and Spanish datasets were doubly-annotated in order to compute inter-annotator agreement (IAA). The overall target identification IAA was 82.4% F_1 for English, 81.6% for Danish, and 80.0% for Spanish. This is lower than a similar monolingual annotation experiment recently reporting target identification IAA at 95.3% (Søgaard et al., 2015b). The frame identification IAA scores were also higher in that study, at 84.5% and 78.1% F_1. The drop in agreement seems mostly due to pre-tagging errors caused by erroneous or irrelevant word-to-word translations. The Spanish data has the lowest agreement score.

We compute test-retest reliability of our annotations as the correlation coefficient (Pearson's ρ) between the two annotations. In Cronbach's α internal consistency table, the cut-off for acceptable reliability is 0.7. While there is certainly noise in our annotations, these are still consistently above the Cronbach cut-off. Also, we evaluate our models across 18 datasets, covering nine different languages with two domains each; although for readability, we combine the Wiktionary and Twitter datasets for each language below.

[2] `http://babelnet.org/`

	Language		
	EN	DA	ES
Twitter and Wikipedia			
TARGET	82.4	81.6	80.0
FRAME	73.5	72.3	60.8
ARGUMENT	70.7	55.0	83.5
Test-retest reliability	74.4	78.6	71.8
Twitter			
TARGET	79.1	80.7	80.5
FRAME	68.8	72.3	58.6
ARGUMENT	70.0	86.2	57.5
Test-retest reliability	71.0	78.7	73.1

Table 1: Inter-annotator agreement (F_1 in %)

The relatively low reliability compared to previous annotation efforts is due to the cross-lingual pre-annotation step, which was necessary to make annotation feasible. All languages, including English, have been pre-annotated using BABELNET. We expect annotators to only assign frames when meaningful frames can be assigned, so the main source of error is that the pre-annotation may exclude valid frames. Hence, we will not only report F_1-scores in our evaluations, but also precision, since recall may be misleading, penalizing for frames that could not be chosen by the annotators.

3 Frame semantic parsing

3.1 Target identification

Following Das et al. (2014), we use part-of-speech heuristics to identify the words that evoke frames (target words). Frame-evoking words typically belong to a narrow range of part of speech. Therefore, we only consider words as target candidates

when tagged with one of the top k part-of-speech tags most commonly seen as targets in the training set. The k parameter is optimized to maximize F_1 on our development language, Spanish, where we found $k = 7$.[3] Surviving candidates are then translated into English by mapping the words into multi-lingual BABELNET synsets, which represent sets of words with similar meaning across languages. All English words in the BABELNET synsets are considered possible translations. If any of the translations are potential targets in FRAMENET 1.5, the current word is identified as a frame-evoking word.

3.2 Frame identification

A target word is, on average, ambiguous between three frames. We use a multinomial log-linear classifier[4] (with default parameters) to decide which of the possible frames evoked by the target word that fits the context best. Our feature representation replicates that of Das et al. (2014) as far as possible, considering the multilingual setting where lexical features cannot be directly used. To compensate for the lack of lexical features, we introduce two groups of language-independent features that rely on multilingual word embeddings. One feature group uses the embedding of the target word directly, while the other is based on distance measures between the target word and the set of English words used as targets for a possible frame. We measure the minimum and mean distance (in embedding space) from the target word to the set of English target words, as well as the distances to each word individually.

Several of the features in the original representation are built on top of automatic POS annotation and syntactic parses. We use the Universal Dependencies v1.1 treebanks for the languages in our data to train part-of-speech taggers (TREETAGGER[5]) and a dependency parser (TURBOPARSER[6]) to generate the syntactic features. In contrast to Das et al. (2014), we use dependency subtrees instead of spans.

[3]The white-listed POS are nouns, verbs, adjectives, proper nouns, adverbs, and determiners.

[4]http://hunch.net/~vw/

[5]http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[6]http://www.cs.cmu.edu/~ark/TurboParser/

3.3 Argument identification

A frame contains a number of named arguments that may or may not be expressed in a given sentence. Argument identification is concerned with assigning frame arguments to spans of words in the sentence. While this task can benefit from information on the joint assignment of arguments, Das et al. (2014) report only an improvement of less than 1% in F_1 using beam search to approximate a global optimal configuration for argument identification. To simplify our system, we take all argument-identification decisions independently. We use a single classifier for argument identification, computing the most probable argument for each frame element. Each word index is associated with a span by the transitive closure of its syntactic dependencies (i.e. subtree). Our greedy approach to argument identification thus amounts to scoring the $n + 1$ possible realisations of an argument for an n-length sentence (i.e. subtrees plus the empty argument), selecting the highest scoring subtree for each argument type allowed by the frame.

As the training data contains very few examples of each frame or role (e.g., *Buyer* in the frame COMMERCE_SCENARIO), we enable sharing of features for frame arguments that have the same name. The assumption is that arguments with identical names have similar semantic properties across frames; that is the argument *Perpetrator*, for example, is similar for the frames ARSON and THEFT.

The scores are the confidences of a binary classifier trained on <*frame*, *argument*, *subtree*> tuples. Positive examples are the observed arguments. We use the remaining n incorrect subtrees for a given <*frame*, *argument*> pair to generate negative training examples . A single binary classification model is trained for the whole data set.

As with frame identification, our features are similar to those of Das et al. (2014), with a few exceptions and additions. We use dependency subtrees instead of spans and replace all lexical features (which do not transfer cross-lingually) with features based on the interlingual word embeddings from Søgaard et al. (2015a). We use the embeddings to find the 20 most similar words in the training data and use these words to generate lexical features that matched the source-language training data. Each feature is weighted by its cosine similarity with the target-language word.

Target identification		BG	DA	DE	EL	EN	ES	FR	IT	SV	Avg.
F_1	SYSTEM	**85.5**	**73.6**	**58.4**	**52.9**	**80.2**	**89.1**	**66.1**	**69.0**	**72.8**	**72.0**
	BASELINE	44.0	56.8	27.2	46.1	78.8	45.9	42.8	47.7	41.4	47.9
Precision	SYSTEM	**89.2**	**70.9**	**66.2**	36.4	**96.3**	**84.9**	51.8	53.4	**63.4**	**67.0**
	BASELINE	56.8	65.0	48.7	**43.2**	88.0	75.2	**55.0**	**55.3**	47.3	59.4

Frame identification		BG	DA	DE	EL	EN	ES	FR	IT	SV	Avg.
F_1	SYSTEM	**66.6**	**59.0**	49.0	**58.3**	37.0	**36.9**	**27.4**	**40.2**	49.5	**47.1**
	BASELINE	19.3	14.1	08.5	12.6	**48.8**	08.2	10.4	15.0	10.1	16.3
	MFS	65.3	54.3	**53.0**	56.2	38.0	34.4	25.5	33.0	**55.3**	46.1
Precision	SYSTEM	**72.8**	**64.7**	**57.9**	**67.1**	49.3	45.6	**36.9**	**47.1**	**65.5**	**56.3**
	BASELINE	37.0	26.4	19.0	27.9	**62.4**	15.7	22.0	25.5	28.3	29.7
	MFS	67.7	59.4	57.4	60.1	46.1	42.3	33.4	41.5	61.5	52.2

Argument identification		BG	DA	DE	EL	EN	ES	FR	IT	SV	Avg.
F_1	SYSTEM	**40.8**	**36.0**	**28.5**	**39.3**	25.3	**19.8**	**18.0**	**26.3**	**28.7**	**29.2**
	BASELINE	26.5	10.5	06.2	09.7	**69.6**	04.6	08.6	14.6	08.6	17.7
Precision	SYSTEM	**39.6**	**33.3**	**26.3**	**36.7**	24.0	**18.1**	**16.8**	**24.8**	**26.4**	**27.3**
	BASELINE	16.2	09.5	05.7	08.8	**66.8**	04.1	08.1	13.8	08.0	16.8

Table 2: Frame semantic parsing results (precision and F_1 in %)

Baseline Our approach to multi-lingual frame semantics parsing extends Das et al. (2014) to cross-lingual learning using the interlingual embeddings from Søgaard et al. (2015a). Our baseline is a more direct application of the SEMAFOR system[7] (Das et al., 2014), translating target language text to English using word-to-word translations and projecting annotation back. For word-to-word translation we use Wiktionary bilingual dictionaries (Ács, 2014), and we use frequency counts from UKWAC[8] to disambiguate words with multiple translations, preferring the most common one. The baseline and our system both use the training data supplied with FRAMENET for learning.

4 Results

Consider first the target identification results in Table 2. We observe that using BABELNET and our re-implementation of Das et al. (2014) performs considerably better than running SEMAFOR on Wiktionary word-by-word translations.

Our frame identification results are also presented in Table 2. Our system is better in six out of nine cases, whereas the most frequent sense baseline is best in two. It is unsurprising that English fares best in this setup, because it does not undergo the word-to-word translation of the other data sets.

Argument identification is a harder task, and scores are generally lower; see the lower part of Table 2. Also, note that errors percolate: If we do not identify a target, or mislabel a frame, we can no longer retrieve the correct arguments. Nevertheless, we observe that we are better than running SEMAFOR on word-by-word translations in eight out of nine languages—all, except English.

Generally, we obtain error reductions over our baseline of 46% for target identification, 37% for frame identification, and 14% for argument identification. For English, we are only 2% (absolute) below IAA for target identification, but about 40% below IAA for frame and argument identification. For Danish, the gap is smaller.

If we compare performance on Wikipedia and Twitter datasets, we see that target identification and frame identification scores are generally higher for Wikipedia, while argument identification scores are higher for Twitter. While Wikipedia is generally more similar to the

[7] http://www.ark.cs.cmu.edu/SEMAFOR/
[8] http://wacky.sslmit.unibo.it/

newswire/balanced corpus in FRAMENET 1.5, the sentence length is shorter in tweets, making it easier to identify the correct arguments.

5 Conclusions

We presented a multi-lingual frame-annotated corpus covering nine languages in two domains. With this corpus we performed experiments to predict target, frame and argument identification, outperforming a word-to-word translated baseline running on SEMAFOR. Our approach is a delexicalized version of Das et al. (2014) with a simpler decoding strategy and, crucially, using multilingual word embeddings to achieve any-language frame-semantic parsing. Over a baseline of using SEMAFOR with word-to-word translations, we obtain error reductions of 46% for target identification, 37% for frame identification, and 14% for argument identification.

References

Judit Ács. 2014. Pivot-based multilingual dictionary building using wiktionary. In *LREC*.

Dipanjan Das, Desai Chen, Andre Martins, Nathan Schneider, and Noah Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *ACL*.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *NODALIDA*.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015a. Inverted indexing for cross-lingual nlp. In *ACL*.

Anders Søgaard, Barbara Plank, and Hector Martinez Alonso. 2015b. Using frame semantics for knowledge extraction from twitter. In *AAAI*.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *ACL*.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient inference and structured learning for semantic role labeling. *TACL*.

Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *ACL*.