Automatic Detection of Light Verb Constructions in Different Languages With Syntax-based Methods

István Nagy T.
Department of Informatics,
University of Szeged
nistvan@inf.u-szeged.hu

Veronika Vincze
Hungarian Academy of Sciences,
Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

1 Introduction

Here, our goal is to automatically identify all occurrences of light verb constructions (LVCs) in raw text in four different languages. In order to do this, we use the 4FX parallel corpus (Rácz et al., 2014), where LVCs were manually annotated in four different languages such as English, German, Spanish and Hungarian. Basically, we apply an English-based machine learning method to automatically detect LVCs and we also introduce how to adapt this method to the other languages. Moreover, we also examine how data from other languages can be exploited in supervised LVC detection for a given language, therefore an implementation of the language-independent representation of LVCs is required.

2 Language Adaptation

We applied an adaptation technique similar to domain adaptation to investigate how data from other languages can influence the performance on LVC detection. Domain adaptation is most successful when we only have a limited set of annotated data from one domain, but there are a plenty of data available in another domain, and we can apply domain adaptation methods to achieve better results on the target domain. We treated the different parts of the 4FX corpus with different languages as different domains. In most cases, domain adaptation can enhance the results if we only have a limited amount of annotated target data and here, we tested this approach on data from different languages. We used one language as the source and we selected another language as the target.

3 Standardized Feature Representation on Different Languages

For the automatic classification of the candidate LVCs a machine learning-based approach (Vincze et al., 2013) was utilized, which was successfully applied to English and Hungarian. This method follows a two-step approach: first, it selects LVC candidates from texts with the help of syntactic information and then it classifies them as genuine LVCs or not, based on a rich feature set with statistical, lexical, morphological, syntactic and orthographic features. In order to detect Spanish and German LVCs in texts, we also implemented this feature set in both new languages and we defined some new language specific features too.

Furthermore, a standardized representation of the feature set is also required, as we also would like to investigate how the different languages can influence each other. Therefore, the same features in different languages were associated with each other: the most typical light verbs in each language were paired with their equivalents in the other languages, forming quadruples such as *take - nehmen - tomar - vesz*. Moreover, syntactic relations and morphological features were also standardized across languages.

The feature set includes language-independent and language-specific features as well. Mainly the language-independent features aim to grab general features of LVCs while language-specific features can be applied due to the different grammatical characteristics of the four languages.

4 Experiments

The standardized feature representation of LVCs on different languages allowed us to focus on the portability of models trained on different parts of the 4FX corpus. We were also able to investigate the effect of the language adaptation techniques so as to reduce the gap between the language pairs.

First, we trained the J48 decision tree classifier with the language-independent feature set and evaluated in a 10-fold cross-validation scheme at the candidate level for each language in the 4FX corpus.

As a baseline, a context-free dictionary lookup method was applied in the four languages. As a dictionary, manually compiled lists were used in each language. In the case of English, manually annotated LVCs were collected from the English part of the SzegedParalellFX corpus (Vincze, 2012), while the filtered version of the German PP-Verb Collocations list (Krenn, 2008) was applied in German. The filtered version of the Spanish verb-noun lexical function dictionary (Kolesnikova and Gelbukh, 2010) was utilized in Spanish and the manually annotated LVCs from the Hungarian part of the above mentioned parallel corpus were collected for Hungarian.

To compare the different languages, a pure cross-language setting was utilized, where our model was trained on the source language and evaluated on the target (i.e. no labeled target language datasets were used for training); e.g. we trained the model on the English part of corpus and tested it on the Hungarian part. Table 1 shows the results of cross language experiments, where the bold numbers show the results of the in-language 10 fold cross-validation and the differences relative to the dictionary lookup methods are also presented.

Test - Train	Dict.	Cross	Diff.
En - En	31.92	**65.35**	+33.43
En - De	31.92	46.31	+14.39
En - Es	31.92	32.34	+0.42
En - Hu	31.92	40.18	+8.26
De - De	13.71	**50.64**	+36.93
De - En	13.71	24.12	+10.41
De - Es	13.71	17.64	+3.93
De - Hu	13.71	10.06	-3.65
Es - Es	40.28	**52.90**	+12.62
Es - En	40.28	32.02	-8.26
Es - De	40.28	31.25	+9.03
Es - Hu	40.28	38.98	-1.3
Hu - Hu	35.34	**64.72**	+29.38
Hu - En	35.34	49.41	+14.07
Hu - De	35.34	48.24	+12.9
Hu - Es	35.34	29.19	-6.15

Table 1: Results of the cross language setting in terms of F-score on the 4FX corpus.

Furthermore, a very simple approach was used for language adaptation: we applied 10 fold cross validation and for each fold, we used 10% of the target language as test and the other 90% was held out for training. The source language training dataset (in columns) was extended with the instances from the target language training dataset (in rows). Table 2 lists the results for the language adaptation, relative to the indomain settings.

	EN		**DE**		**ES**		**HU**	
EN	**65.35**	–	65.38	+0.03	65.69	+0.34	65.58	+0.23
DE	51.17	+0.53	**50.64**	–	51.23	+0.59	50.74	+0.1
ES	51.86	-1.04	53.54	+0.64	**52.90**	–	53.09	+0.19
HU	65.25	+0.53	64.58	-0.14	64.69	-0.03	**64.72**	–

Table 2: Results of the language adaptation setting in terms of F-score on the 4FX corpus.

5 Discussion

Our results reveal that the results of our cross-language experiments mostly exceeded those of the dictionary lookup method. This highlights the fact that a machine learning-based model trained on a different language can be more effective than a target language dictionary lookup method.

Moreover, the cross-language training by itself did not prove sufficient in many cases. Therefore, the inclusion of annotated target language data into the training dataset was necessary to reduce the gap among languages. By adding some target language data to the training dataset, it is possible to achieve similar or even better results compared to those of in-domain settings. Thus, a simple language adaptation technique may prove useful in syntax-based cross-lingual LVC detection.

References

Olga Kolesnikova and Alexander Gelbukh. 2010. Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish. In *Advances in Soft Computing*, pages 196–207. Springer.

Brigitte Krenn. 2008. Description of Evaluation Resource – German PP-verb data. In *Proceedings of MWE 2008*, pages 7–10, Marrakech, Morocco, June.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. In *Proceedings LREC'14*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013. Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. *Proceedings of ACL-2013: Short Papers, Sofia. ACL.*

Veronika Vincze. 2012. Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In *Proceedings of LREC 2012*, Istanbul, Turkey.