WG3

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

István Nagy T. and Veronika Vincze
University of Szeged, Hungary
{nistvan,vinczev}@inf.u-szeged.hu

Motivation

- Automatically identify all occurrences of LVCs in raw texts in four different languages
- English-based machine learning method [2] was adapted to other languages
- We examine how data from other languages can be exploited in supervised LVC detection
- Language-independent representation of LVCs is implemented

Language Adaptation

- Similar to domain adaptation
- The different parts of the 4FX corpus with different languages treated as as different domains
- One language as the source and another language as the target
- Domain adaptation can enhance the results if we only have a limited amount of annotated target data
- A simple approach was used for language adaptation: the source language training dataset was extended with instances from the target language training dataset

	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	+0.10
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	-

Experiments

- J48 decision tree classifier with the language-independent feature set and evaluated in a 10-fold cross-validation
- A context-free dictionary lookup method was applied as baseline in the four languages
- To compare the different languages, a pure cross-language setting was utilized

References

1. 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).
2. 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.

The 4FX corpus

- Texts from JRC-Acquis
- Legal domain
- English, German, Spanish and Hungarian
- Approximately 100K tokens for each language
- Manual annotation for LVCs [1]

Standardized feature representation

- Investigate how the different languages can influence each other
- Standardized representation of the feature set is also required
- 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

take - nehmen - tomar – vesz

- Syntactic relations and morphological features were also standardized across languages
- The language-independent features aim to grab general features
- Language-specific features can be applied due to the different grammatical characteristics of the four languages

Results of the cross language setting in term of F-score

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