USzeged: Identifying Verbal Multiword Expressions with POS Tagging and Parsing Techniques

Katalin Ilona Simkó[1], **Viktória Kovács**[2] **and Veronika Vincze**[3]

[1,3]University of Szeged, Institute of Informatics
[1,2]University of Szeged, Department of General Linguistics
[3]MTA-SZTE Research Group on Artificial Intelligence
[1]simko@hung.u-szeged.hu
[2]viktoria.kovacs12@gmail.com
[3]vinczev@inf.u-szeged.hu

Abstract

The paper describes our system submitted for the Workshop on PARSEME's Shared Task on automatic identification of verbal multiword expressions . It uses POS tagging and dependency parsing to identify single- and multi-token verbal MWEs in text. Our system is language-independent and competed on nine of the eighteen languages. Our paper describes how our system works and gives its error analysis for the languages it was submitted for.

1 Introduction

In our paper, we give a description of the USzeged team's system for the shared task on automatic identification of verbal multiword expressions. We used POS tagging and dependency parsing to identify the verbal MWEs in the text. Our system is language-independent, but relies on POS tagged, dependency analyzed training data. We submitted results for nine out of the eighteen languages, but could be extended to any language if provided with POS tagging and dependency analysis of the training database.

In the paper, we first describe how the system works in detail, then show the results achieved in the shared task on the nine languages with both POS tagging and dependency analysis, last we give an error analysis of our output.

2 Shared task

Our system was built for the shared task on automatic identification of verbal multiword expressions[1] organized as part of the 2017 MWE workshop.

The shared task's aim is to identify verbal MWEs in multiple languages. In total, 18 languages are covered that were annotated using guidelines taking universal and language-specific phenomena into account.

The guideline identifies five different types of verbal MWEs: idioms (ID), light verb constructions (LVC), verb-particle constructions (VPC), inherently reflexive verbs (IRefIV) and other. Their identification in NLP is difficult because they are often discontinuous and non-compositional, the categories are heterogeneous and the structures show high syntactic variability.

Our team created the Hungarian shared task database and VMWE annotation. Our system is mostly based on our experiences with the Hungarian data in this annotation phase.

3 System description

Our system works through the connection of MWEs and parsing, an approach described by many sources (Constant and Nivre, 2016; Nasr et al., 2015; Candito and Constant, 2014; Green et al., 2011; Waszczuk et al., 2016; Wehrli et al., 2010; Green et al., 2013) and is one the basic ideas behind the work done by the PARSEME group [2].

The idea for our system is directly based on the work described in Vincze et al. (2013) to use dependency parsing to find MWEs. As a high number of the languages of the shared task are morphologically rich and have free word order, therefore syntactically flexible MWEs might not be adjacent, this approach seems a better fit for the task than sequence labeling or similar strategies.

The system of that paper uses dependency relations specific to syntactic relation and MWE type, for example light verb constructions that are made up of a verb-object relation syntactically, get the

[1]http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_05_MWE_2017___lb__EACL__rb__&subpage=CONF_40_Shared_Task

[2]http://typo.uni-konstanz.de/parseme/

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 48–53,
Valencia, Spain, April 4.

label OBJ-LVC in the merged annotation.

In contrast, our system uses only the MWE type as a merged dependency label and it also applies to single-token MWEs. As multiple languages had single-token MWEs as well as multi-token ones dealt with in dependency parsing, we expanded the approach using POS tagging.

MWEs have specific morphological, syntactic and semantic properties. Our approach treats multi-token MWEs on the level of syntax – similarly to the `MWE` dependency relation in the Universal Dependency grammar (Nivre, 2015) – and single-token MWEs on the level of morphology.

Our system works in four steps, and the main MWE identification happens within POS tagging and dependency parsing of the text. Our system relies on the POS tagging and dependency annotations provided by the organizers of the shared task in the companion CoNLL files and the verbal MWE annotation of the texts and is completely language-independent given those inputs.

In the first step, we prepared the training file from the above mentioned inputs. We merged the training MWE annotation into its dependency annotation for single and multi-token MWEs separately. The single-token MWEs POS tag got replaced with their MWE type, while for the multi-token MWEs the dependency graphs' label changed: the label of the token lower in the tree was replaced with a label with the MWE type.

Figures 1-3 show the single-token MWE's change in POS tag and multi-token MWE dependency relabeling for VPCs and LVCs in a Hungarian example.

For multi-token MWEs our approach is based on our theory that the lower MWE element will be directly connected to the other MWE element(s). We do not change the structure of the dependency relations in the tree, but change the dependency label of the lower MWE element to the MWE type, therefore making the MWE element retraceable from the dependency annotation of the sentence. For example *lát* and *el* in Example 2 make up a VPC, so the dependency relation label of the lower element, *el* changes from the general syntactic label **PREVERB** to the MWE label **VPC**, with this **VPC** label now connecting the two elements of the MWE.

For MWEs of more than two tokens, the conversion replaces the dependency labels of all MWE elements below the highest one. In example 4, the highest element of the idiom *az első követ veti* ("casts the first stone") is the verb, *vetette* (cast.Sg3.Past). All other elements' dependency labels are changed to **ID**.

The second step is training the parser: we used the Bohnet parser (Bohnet, 2010) for both POS tagging and dependency parsing. For the single-token MWEs, we trained the Bohnet parser's POS tagger module on the MWE-merged corpora and its dependency parser for the multi-token MWEs. The parser would treat the MWE POS tags and dependency labels as any other POS tag and dependency label.

We did the same for each language and created POS tagging and dependency parsing models capable of identifying MWEs for them. In the case of some of the languages in the shared task, we had to omit sentences from the training data that were overly long (spanning over 500 tokens in some cases) and caused errors in training.

Third, we ran the POS tagging and dependency parsing models of each language on their respective test corpora. The output contains the MWE POS tags and dependency labels used in that language as well as the standard POS and syntactic ones.

The fourth and last step is to extract the MWE tags and labels from the output of the POS tagger and the dependency parser. The MWE POS tagged words are annotated as single-token MWEs of the type of their POS tag. From the MWE dependency labels, we annotate the words connected by the MWE label as making up a multi-token MWE of that type.

4 Results

We submitted our system for all languages in the shared task with provided dependency analysis and POS tagging. POS tagging was needed for the single-token MWEs frequent in some languages, while we used dependency analysis in identifying multi-token MWEs. We attempted to use just the POS tagging component of our system on the languages that only had POS tagging available to give partial results (i.e. identifying only single-token MWEs), but we found that these languages incidentally had no or very few single-token MWEs, therefore not providing adequate training data.

Our results on the nine languages are in Table 1. Our system was submitted for German, Greek, Spanish, French, Hungarian, Italian, Polish, Por-

bekezdés	NOUN	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
bekezdés	VPC	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
határozathozatal	NOUN	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none
határozathozatal	LVC	SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none

Figure 1: Adding the VPC and LVC single-token MWE POS tags to *bekezdés* (lit. in+starting, "paragraph") and *határozathozatal* (lit. decision+bringing, "decision-making").

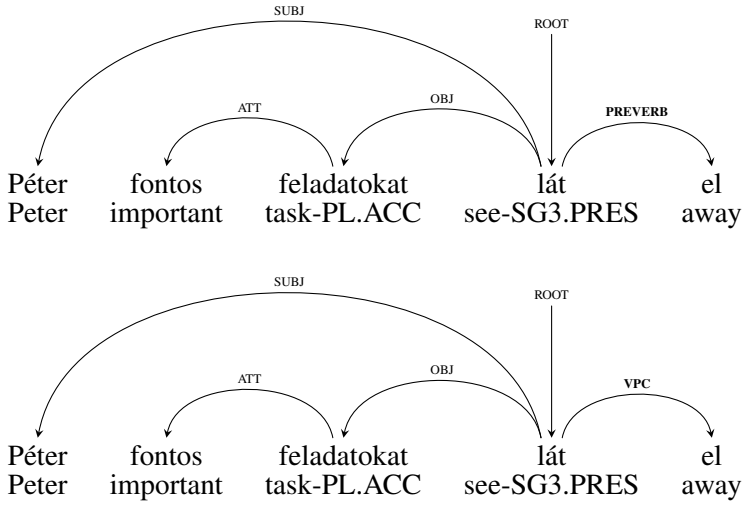

Figure 2: Adding the VPC multi-token MWEs label to the dependency graph in the sentence *Peter takes care of important tasks*.

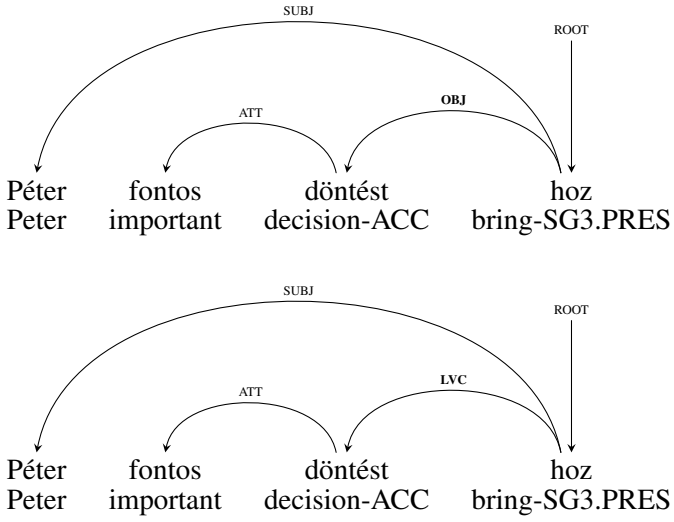

Figure 3: Adding the LVC multi-token MWE label to the dependency graph in the sentence *Peter makes an important decision*.

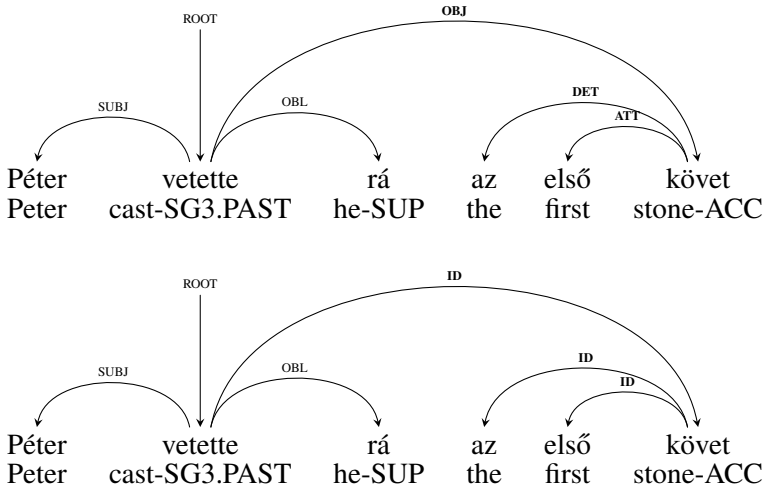

Figure 4: Adding the ID multi-token MWE label to the dependency graph in the sentence *Peter cast the first stone on him*.

tuguese, and Swedish.

The F-scores show great differences between languages, but so did they for the other systems entered. Compared to the other, mostly closed track systems, the USzeged system ranked close to or at the top on German, Hungarian, and Swedish. For the other languages (except for Polish and Portuguese, where ours is the worst performing system), we ranked in the mid-range. These results are related to the way our system works and the verbal MWE types frequent in the languages.

5 Error analysis

After receiving the gold annotation for the test corpora, we investigated the strengths and weaknesses of our system.

The shared task data was annotated for five types of verbal MWEs: light verb constructions, verb-particle constructions, inherently reflexive verbs, idioms, and "other".

Our error analysis showed that our system performs by far best on the verb-particle construction category, correctly identifying around 60% of VPCs, but only about 40% of other types. Verb-particle constructions are most likely to have a syntactic relationship between the MWE elements, which would support why our system is good at identifying them.

German, Hungarian, and Swedish were also the languages with the highest proportions of the VPC type of verbal MWEs in the shared task, which also correlates with why our system performed best on them. Romance languages contain almost no VPCs and the remaining ones have much less also. In this way, our achieved results seem to be dependent on the type of verbal MWEs frequent in that language because of the inherent characteristics of the system.

For French and Italian, our system also performed worse on IRefIVs. Generally, we had some trouble identifying longer IDs and LVCs and MWEs including prepositions. A further source of error was when there was no syntactic edge in between members of a specific MWE, for instance, in German, the copula *sein* "be" was often indirectly connected to the other words of the MWE (e.g. *im Rennen sein* "to compete"), hence our method was not able to recognize it as part of the MWE. We plan to revise our system to not only relabel dependency relations, but also restructure a tree in an attempt to deal with these issues.

6 Conclusions

In our paper, we described the USzeged verbal MWE identifying tool developed for the PARSEME Shared Task. Our system merged the MWE annotation with the POS tagging and dependency annotation of the text and used a standard POS tagger and dependency parser to identify verbal MWEs in texts. The system is language-independent given those inputs, but the overall results it achieves seem to rely on the type of verbal MWEs frequent in the given language.

	System	P-MWE	R-MWE	F-MWE	P-token	R-token	F-token
DE	BEST, USZEGED	0.5154	0.3340	0.4053	0.6592	0.3468	0.4545
	LAST	0.3652	0.1300	0.1917	0.6716	0.1793	0.2830
EL	BEST	0.3612	0.4500	0.4007	0.4635	0.4742	0.4688
	USZEGED	0.3084	0.3300	0.3188	0.4451	0.3757	0.4075
	LAST	0.4286	0.2520	0.3174	0.5616	0.2953	0.3871
ES	BEST	0.6122	0.5400	0.5739	0.6574	0.5252	0.5839
	USZEGED	0.2575	0.5000	0.3399	0.3635	0.5629	0.4418
	LAST	0.6447	0.1960	0.3006	0.7233	0.1967	0.3093
FR	BEST	0.6147	0.4340	0.5088	0.8088	0.4964	0.6152
	USZEGED	0.0639	0.0520	0.0573	0.5218	0.2482	0.3364
	LAST	0.8056	0.0580	0.1082	0.8194	0.0532	0.1000
HU	BEST, USZEGED	0.7936	0.6934	0.7401	0.8057	0.6317	0.7081
	LAST	0.8029	0.5471	0.6508	0.8208	0.5015	0.6226
IT	BEST	0.5354	0.3180	0.3990	0.6134	0.3378	0.4357
	USZEGED	0.1503	0.1560	0.1531	0.4054	0.3064	0.3490
	LAST	0.6125	0.0980	0.1690	0.6837	0.1053	0.1824
PL	BEST	0.7798	0.6020	0.6795	0.8742	0.6228	0.7274
	LAST, USZEGED	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PT	BEST	0.7543	0.6080	0.6733	0.8005	0.6370	0.7094
	LAST, USZEGED	0.0129	0.0080	0.0099	0.6837	0.1987	0.3079
SV	BEST	0.4860	0.2203	0.3032	0.5253	0.2249	0.3149
	USZEGED	0.2482	0.2966	0.2703	0.2961	0.3294	0.3119
	LAST	0.5758	0.1610	0.2517	0.6538	0.1677	0.2669

Table 1: Best, last and USzeged systems' results for the languages ranked by per-token F-scores.

References

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Baltimore, Maryland, June. Association for Computational Linguistics.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 161–171, Berlin, Germany, August. Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, 39(1):195–227.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1116–1126, Beijing, China, July. Association for Computational Linguistics.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Veronika Vincze, János Zsibrita, and István Nagy T. 2013. Dependency parsing for identifying hungarian light verb constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. 2016. Promoting multiword expressions in A* TAG parsing. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 429–439.

Eric Wehrli, Violeta Seretan, and Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 27–35, Beijing, China, August. Association for Computational Linguistics.