Finding English Equivalents of Hungarian Light Verb Constructions WG3

István Nagy T.[1], **Veronika Vincze**[1,2]
[1]Institute of Informatics, University of Szeged
Árpád tér 2., 6720 Szeged, Hungary
nistvan@inf.u-szeged.hu
[2]MTA-SZTE Research Group on Artificial Intelligence
Tisza Lajos krt. 103., 6720 Szeged, Hungary
vinczev@inf.u-szeged.hu

In this study we present our machine learning-based method to automatically identify the English equivalents of Hungarian light verb constructions (LVCs) based on a Hungarian – English parallel corpus. The main difficulties of the automatic identification of LVC translations lie in the fact that the meaning of LVCs can only partially be computed on the basis of the meanings of their parts and the way they are related to each other (semi-compositionality). Thus, the result of translating their parts literally can hardly be considered as the proper translation of the original expression.

Earlier studies like Seretan (2015) evaluated the efficacy of translating multiword expressions by machine translation approaches and they showed that the automatic translation of such expressions can be complicated in the case of many languages. For this reason, various studies have been conducted that aimed at the translation of multiword expressions, mostly with the purpose of supporting machine translation systems (Monti et al., 2013). Most of these methods (Monti et al., 2015; Wehrli and Villavicencio, 2015) can automatically identify multiword expressions in various languages by using automatic approaches, then they provide different solutions for the selection of possible translation pairs. Here, our approach is based on similar principles and we present our machine learning based methods to automatically identify the English equivalents of Hungarian light verb constructions in the SzegedParalellFX corpus (Vincze, 2012), which contains manually annotated LVCs in both languages.

First, we made use of the gold standard annotation of Hungarian LVCs, and generated their potential English equivalents. We supposed that the same sentence alignment unit will contain the translational equivalent of the LVC, so we applied a syntax-based method (Vincze et al., 2013) to create all the possible LVCs in the English part of the text. From the candidates, a linguist selected the correct translation for the original LVC, paying attention to the fact that it was required to be another LVC. Thus, if the Hungarian LVC *döntést hoz* decision-ACC bring was translated with a verbal synonym (eg. *decide* instead of *make a decision*), it was not accepted.

Table 1: Results of the baseline method and machine learning-based approach

Approach	**Precision**	**Recall**	**F-measure**
Baseline	73.68	15.69	25.88
Decision tree	47.63	54.93	50.81

Second, we wanted to see how potential translation pairs can be automatically identified, that is, how correct translations can be selected by using a machine learning-based approach. We used the J48 classifier of the WEKA package (Hall et al., 2009), which implements the decision trees algorithm C4.5 (Quinlan, 1993) with a feature set optimized for English–Hungarian LVC detection (Vincze et al., 2013) and we applied 10-fold cross validation on the SzegedParallelFX corpus.

As a baseline method, we applied a context-free dictionary lookup method. We treated the potential translation pairs as correct translations when the two nominal components proved to be translational equivalents in a Hungarian-English dictionary. Table 1 lists the results of both the baseline dictionary lookup method and our machine learning approach.

As the negative examples were overrepresented in the training set, we gave extra weight to the positive examples during the training process. For finding the optimal F-measure, we investigated the efficiency of the machine learning-based method with different weights. Results are shown in Figure 1.

As can be seen, when the weight of positive elements was increased, the precision of the machine learning approach was decreased and recall was increased. In our experiments, adding triple weight to positive examples resulted in the highest F-score. Results also show that the quality of the automatic dictionary can be modified by weights depending of the end application. If the aim is to create an accurate dictionary, then low weight should be assigned to positive samples as precision will increase. If the aim is to get many possible translation pairs, high weight should be applied during machine learning as recall will increase. However, due to the difficulty of the task, manual validation of the automatic dictionary is necessary in all cases.

As a result of error analysis, we found that the most difficult task was when both the English and Hungarian translation units contained LVCs but they but did not correspond to each other. For example:

Háromévi várakozás után William
of.three.years waiting after William
Prichard kapitány, az Antilop
Prichard captain, the Antelope
gazdája, ki a déli
owner-3SGPOSS, who the Southern
vizekre volt indulóban, előnyös
water-PL-SUB was leave-INE, advantageous
ajánlatot tett *nekem, és én*
offer-ACC make-PAST-3SG I-DAT, and I
elfogadtam.
accept-PAST-1SG-OBJ.
'After three years of expectation, captain William Prichard, the master of Antelope, who was about to leave to the South Sea, made me an advantageous offer and I accepted it.'

After three years' expectation that things would mend, I accepted an advantageous offer from Captain William Prichard, master of the Antelope, who was ***making a voyage*** *to the South Sea.*

In this case the expressions *ajánlatot tett* "make an offer" and *making a voyage* were identified as pairs, however, this is not correct. It was also problematic for the system to identify LVCs which contain verbs rarely used in LVCs: *(nehéz) életet élnek* (difficult) life-ACC live-3PL - *lead (difficult) lives*.

Our Hungarian–English LVC pairs generated in this way will be made freely available for the community.

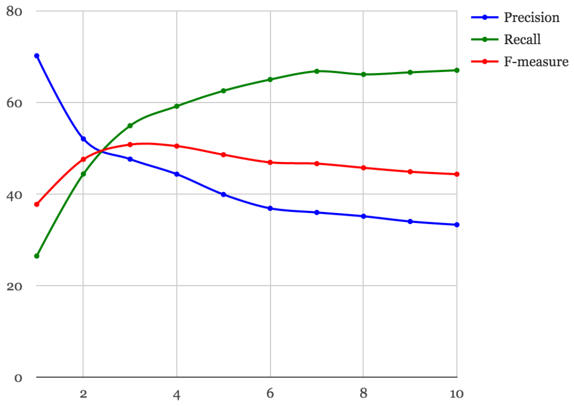

Figure 1: The effect of weights of the positive elements on the efficiency of the machine learning-based approach.

References

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan. 2013. Multi-word units in machine translation and translation technologies.

Johanna Monti, Federico Sangati, and Mihael Arcan. 2015. Multi-word expressions in a parallel bilingual spoken corpus: data annotation and initial identification results. Poster. PARSEME 5th General Meeting.

J. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Violeta Seretan. 2015. Multi-word expressions in user-generated content: How many and how well translated? evidence from a post-editing experiment. In *Proceedings of MUMTTT 2015*, Malaga, Spain.

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013. Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In *Proceedings of ACL 2013*, pages 255–261.

Veronika Vincze. 2012. Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In *Proceedings of LREC-2012*, pages 2381–2388, Istanbul. ELRA.

Eric Wehrli and Aline Villavicencio. 2015. Extraction of Multilingual MWEs from Aligned Corpora. Poster. PARSEME 5th General Meeting.