Extracting meaning by idiomaticity: Description of the HSemID system at CogALex VI (2020)

Jean-Pierre Colson
University of Louvain
Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Abstract

The HSemID system, submitted to the CogALex VI Shared Task is a hybrid system relying mainly on metric clusters measured in large web corpora, complemented by a vector space model using cosine similarity to detect semantic associations. Although the system reached rather weak results for the subcategories of synonyms, antonyms and hypernyms, with some differences from one language to another, it is able to measure general semantic associations (as being random or not-random) with an F1 score close to 0.80. The results strongly suggest that idiomatic constructions play a fundamental role in semantic associations. Further experiments are necessary in order to fine-tune the model to the subcategories of synonyms, antonyms, hypernyms and to explain surprising differences across languages.

1 Introduction

This paper is a system description of *HSemID (Hybrid Semantic extraction based on IDiomatic associations)*, presented at CogALex VI. Contrary to most models dedicated to the extraction of semantic associations, *HSemID* is based on a similar model developed for the extraction of multiword expressions, *HMSid*, presented at the Parseme 1.2. workshop of the Coling 2020 conference. From a theoretical point of view, we wished to explore the link between general meaning associations and associations based on idiomaticity, in the general sense of multiword expressions (MWEs). For instance, *beans* may display a general meaning association with food (as many of them are edible) or with coffee, but there is an idiomatic association between *spill* and *beans* because of the idiom *spill the beans* (reveal a secret). Thus, general meaning associations are mainly extralinguistic and cultural, whereas idiomatic associations are mainly intralinguistic, as they are just valid for one specific language, although similar associations may exist in other languages because they are cognate or have influenced each other.

The implicit link between semantics and idiomaticity has already been mentioned in the literature. Lapesa and Evert (2014) point out that using larger windows with statistical scores yields extraction models that can be adapted from MWEs to semantic associations. According to them, 1st-order models (based on co-occurrence statistics such as the log-likelihood, dice score or t-score) and 2nd-order models (based on similar contexts of use, as in the case of cosine similarity in a vector space model) appear to be redundant on the basis of the first experiments and do not really benefit from a combination of both approaches.

Our model for the extraction of multiword expressions (*HMSid, Hybrid Multi-layer System for the extraction of Idioms*) yielded promising results for French verbal expressions. In the official results of the Parseme 1.2. shared task, our model obtained an F1-score of 67.1, with an F1-score of 36.49 for MWEs that were unseen in the training data; in an adapted version proposed just after the workshop, we

reached an even better F1-score of 71.86 in the closed track, relying only on the training data, with no external resources, and an F1-score for unseen MWEs of 40.15, which makes it by far the best score in the closed track for unseen French MWEs. It should be pointed out that the model used for the extraction of MWEs is corpus-based and derives from metric clusters used in Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999; Colson, 2017; 2018), but does not use any machine learning architecture.

We adapted this model in a deep learning approach for the CogALex VI Shared Task, as described in the following section. From a theoretical point of view, we wished to explore the performance of a model used for MWE extraction, in a related but very different context: the extraction of semantic associations. Although we realize that the main goal of the CogALex VI Shared Task was to improve the extraction of the specific categories of synonyms, antonyms and hypernyms, we did not have enough time to train our model for this subcategory distinction, and were mainly concerned with the identification of a semantic association (being random or non-random) on the basis of idiomatic patterns.

2 Methodology

Our model was tested for the different languages taking part in the CogALex VI Shared Task, using the datasets provided for English (Santus et al., 2015), Chinese (Liu et al., 2019), German (Scheible and Schulte Im Walde, 2014) and Italian (Sucameli and Lenci, 2017).

As suggested by the acronym (*HSemID, Hybrid Semantic extraction based on IDiomatic associations*), our methodology was hybrid, as we used both a vector space model (Turney and Pantel, 2010) and a co-occurrence model based on metric clusters in a general corpus (Colson, 2017; 2018). However, most features of the model were derived from the second part, so that the model mainly relies on co-occurrence and therefore on *idiomatic* meaning, as explained below.

For the vector space model, we measured cosine similarity in the Wikipedia corpora. We relied on the Wiki word vectors[1] and on the Perl implementation of Word2vec, by means of the multiword cosine similarity function[2].

For the metric cluster, we used the *cpr-score* (Colson, 2017; 2018), a simple co-occurrence score based on the proximity of the ngrams in a large corpus. In order to avoid redundancy with the Wikipedia corpora, this score was computed by using other, general-purpose web corpora: the WaCky corpora (Baroni et al., 2009) for English, German and Italian. For Chinese, we compiled our own web corpus by means of the WebBootCat tool provided by the Sketch Engine[3]. As we have only basic knowledge of Chinese, we relied for this purpose on the seed words originally used for compiling the English WaCky corpus. The English seed words were translated into Chinese by Google Translate[4]. All those corpora have a size of about 1.4 billion tokens; for Chinese (Mandarin, simplified spelling), we reached a comparable size by taking into account the number of Chinese words, not the number of Chinese characters (*hans*).

In order to train our model, we implemented a neural network (multi-layer perceptron), relying on most of the default options provided by the Microsoft Cognitive Toolkit (CNTK)[5]. We imported the CNTK library in a python script. Our neural network used minibatches, had an input dimension of just 11 features (for the 4 output classes), 2 hidden layers (dimension: 7), and we used ReLU as an activation function. For the loss, we relied on cross entropy with softmax.

Among the 11 features used for training the model, it should be noted that the vector space approach, represented by the multiple cosine similarity, only played a limited role, as it represented just one of the 11 features to be weighted by the model. The other features were based on the metric clusters. For these, the association score (*cpr-score*) was measured with a narrow window between the grams composing the pairs from the datasets, and with wider windows for a number of linguistic markers favoring semantic associations (typically *or, and, not,* and their equivalents in the different languages). The frequencies of the different grams in the WaCky corpora were also used as input features. All features were smoothened to real figures between 0 and 1. For measuring the average test error during training, we used 80 percent

[1] The Wiki word vectors can be downloaded from http://fasttext.cc/docs/en/pretrained-vectors.html
[2] https://metacpan.org/pod/Word2vec::Word2vec
[3] https://www.sketchengine.eu/
[4] https://translate.google.com
[5] https://www.microsoft.com/en-us/research/product/cognitive-toolkit

of the training data as the trainer, and 20 percent (with the correct labels) as the test data. The average test error when training the model was situated around 20 percent.

3 Results and discussion

Table 1 below displays the official results obtained by HSemID at the CogALex VI Shared Task for the various languages (English, Chinese, German, Italian).

HSemID			
English	**P**	**R**	**F1**
SYN	0.483	0.214	0.297
HYP	0.416	0.366	0.389
ANT	0.313	0.248	0.277
Overall	0.400	0.276	0.320
Chinese			
SYN	0.282	0.328	0.303
HYP	0.610	0.194	0.294
ANT	0.591	0.458	0.516
Overall	0.501	0.331	0.377
German			
SYN	0.374	0.219	0.276
HYP	0.386	0.273	0.320
ANT	0.422	0.281	0.338
Overall	0.395	0.258	0.312
Italian			
SYN	0.418	0.371	0.393
HYP	0.344	0.294	0.317
ANT	0.319	0.201	0.247
Overall	0.365	0.296	0.325

Table 1: Official results obtained with HSemID at the CogALex VI Shared Task

As shown in Table 1, the overall results yielded by HSemID are situated between an F1 of 0.312 and 0.377. Strangely enough, the best result was reached for Chinese, in spite of the fact that we only have basic mastery of Chinese and have assembled our web corpus, as described in the preceding section, without any feedback from native speakers or specialists of the language. It should also be noted that there is some variation as to the category that receives the best F1 score: English and German score best for hypernyms (respectively 0.389 and 0.320), Chinese for antonyms (0.516) and Italian for synonyms (0.393). Our hypothesis for explaining this phenomenon, in spite of the fact that the methodology was

the same for all languages, is that the hybrid approach checked the cosine similarity in the Wikipedia corpus, but the metric cluster in the web corpora; as the word pairs from the dataset contained several technical terms, the presence or absence of those words in the web corpora was often a matter of pure chance, which may have an influence on the final score from one language to another. The fluctuating results for the Chinese dataset are also striking: not only is the overall F1 score for Chinese the best result of the model, but the model reaches surprising scores for Chinese antonyms (P=0.591, R=0.458), although this category is much more problematic for the European languages.

For lack of time, we didn't have the opportunity of fine-tuning our model to the specific subcategories SYN, HYP and ANT, as was the main goal of the CogALex VI Shared Task. As a matter of fact, our objective was to focus the training of the model on the general semantic associations (random or not-random), in the hope that this would also yield acceptable subcategories SYN, HYP and ANT. Obviously, this was not really the case, although high scores for European languages are hard to reach (the best F1 scores for English, German and Italian at the Shared Task are resp. 0.517, 0.500 and 0.477). A closer analysis of the errors produced by our model reveals that too many idiomatic associations of synonyms and antonyms are similar. For instance, *turn right* and *turn left* are equally strong idiomatic associations, and it is unclear how *right* and *left* should be considered as antonyms from this point of view. In the same way, hypernyms are very hard to discriminate from synonyms if we pay attention to their idiomatic associations. A further improvement of our model may therefore consist in a more complex neural network, in which the different contexts for SYN/ANT and SYN/HYP would be specified by additional features.

In spite of these shortcomings, our model reached pretty good scores for the general task of extracting semantic links, which does not appear in the official results but may be computed by means of the evaluation score provided in the training data, which contains the RANDOM category. If we take into consideration the F1 score obtained for the RANDOM label, we obviously get a picture of the general ability of the model to extract strong semantic associations, be they cases of synonymy, hypernymy or anything else (such as metaphors or idiomatic meaning).

For lack of space, Table 2 below just displays the results obtained in English and Chinese by our model, for the RANDOM category. The scores were computed with the official gold dataset and the original evaluation script included in the training data of the shared task.

HSemID			
English	**P**	**R**	**F1**
RANDOM	0.748	0.822	0.783
Chinese			
RANDOM	0.782	0.807	0.794

Table 2: Results obtained by HSemID for the RANDOM category

It should also be reminded that the best F1 score obtained for this task (subtask 1) at the preceding edition of the CogALex Shared Task[6], CogALex V, was 0.790. After sending the official results of the model to the Shared Task, we continued training the model for English with a more complex neural network and we can report an even better English F1 score: 0.802 (with P=0.716 and R=0.911).

In spite of the rather weak results obtained by our model for the elicitation of the subcategories SYN, HYP and ANT at the CogALex VI Shared Task, we therefore come to the conclusion that the HSemID model, relying mainly on the extraction of semantics by means of idiomatic associations, makes it possible to extract general semantic associations, with F1 figures for the RANDOM category that were rarely reached by any experiment carried out within distributional semantics. The results strongly suggest that idiomatic constructions play a key role in semantic associations. Further experiments should improve the scores obtained for synonyms, antonyms and hypernyms, which clearly remains a daunting challenge in the case of European languages.

[6] https://sites.google.com/site/cogalex2016/home/shared-task/results

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press /Addison Wesley, New York.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43: 209–226.

Jean-Pierre Colson. 2017. The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In Ruslan Mitkov (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Springer International Publishing, Cham: 16–28.

Jean-Pierre Colson. 2018. From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin. In Agatha Savary et al. 2018: 41-50.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Hongchao Liu, Emmanuele Chersoni, Natalia Klyueva, Enrico Santus, and Chu-Ren Huang. 2019. Semantic Relata for the Evaluation of Distributional Models in Mandarin Chinese. IEEE access, 7:145705–145713.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: An Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In Proceedings of the ACL Workshop on Linked Data in Linguistics: Resources and Applications.

Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan and Miriam R. L. Petruck (eds.). 2018. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, *Coling 2018*, Santa Fe NM, USA, Association for Computational Linguistics.

Silke Scheible and Sabine Schulte Im Walde. 2014. A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives. In Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing.

Irene Sucameli and Alessandro Lenci. 2017. PARAD-it: Eliciting Italian Paradigmatic Relations with Crowdsourcing. In Proceedings of CLIC.it.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37:141–188.