Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Predicting default and non-default aspectual coding: Impact und Density of information features[1]

Michael Richter **Tariq Yousef**
Universität Leipzig
Natural Language Processing Group
{richter, tariq}@informatik.uni-leipzig.de

Abstract

This paper presents a study on the automatic classification of default and non-default codings for aspect-marked verbs in six Slavic languages and in Latvian. As classifier a Support Vector Machine and as verbal features *Shannon Information* (SI) and *Average Information Content* (IC) have been utilised. In all languages high accuracy of the classification has been achieved. In addition, we found indications for the validity of the *Uniform Information Density principle* within SI and IC.

1 Introduction

The research questions are: can Shannon's theorem be transferred to natural languages, and, in particular, does coding of aspect marked verbs interact with the information that they carry?

Our point of departure is that verbs have a dominant aspect category and that this category can be determined by frequency distributions: default forms will occur more frequently than non-default forms.

2 Aims, Data and Method

The first aim of the study is to test whether default and non-default coding of aspect-marked verbs in the six Slavic languages Bulgarian, Old Church Slavonic, Polish, Slovak, Slovenian, Ukranian and Latvian can be classified by two verbal information features: *Average Information Content* (henceforth 'IC') (Cohen Priva, 2008; Piantadosi et al., 2011, see (1))

$$IC = E(-log_2(P(W = w \mid context))) \quad (1)$$

As contexts, we took bigrams (*lexical surprisal*, Hale, 2001; Levy, 2008; Levy, 2013), to both directions of the target verbs as a study of Richter et al. (2019) disclosed that target verbs convey the highest amount of information in bigram contexts. We took *Shannon Information* (henceforth 'SI', Shannon and Weaver, 1948) as the negative log probability of a target verb form in the corpus.

The aim and the choice of the two information-theory based features are motivated by Shannon's *source coding theorem* (Shannon and Weaver, 1948) on the interaction of information, coding and length of signs. As classifier we employed a Support Vector Machine (SVM) binary classifier with a radial basis function kernel (Joachims, 1998).

The second aim of the study is to test whether the *Uniform Information Density* – hypothesis, (henceforth UIDh; Genzel and Charniak, 2002; Aylett and Turk, 2004; Levy and Jaeger, 2007; Jaeger, 2010), holds within the features IC and SI of the target verbs. In its original form, UIDh is applied to discrete signs: there should neither be extreme peaks nor extreme troughs in the stream of information in order to facilitate language processing. We, however, apply UIDh to two different information values of a *single* sign and hypothesise based on previous research (Celano et al., 2018) that the variances in information density within IC and SI should tend towards zero (Collins, 2014). We utilised *Global Information Density* $\text{UID}_{\text{GLOBAL}}$ (see (3)): id_i is the information density of SI and IC of a single verb form, and μ is the mean of *id*:

$$UID_{GLOBAL} = -E(\sum_{i=1}^{N} id_i - \mu)^2) \quad (3)$$

As data resource we exploited Universal Dependency Treebanks (version 2.3, https://universaldependencies.org) because verbal aspect is encoded

[1] Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number: 357550571.

in these corpora. For each verb, the default and non-default aspect was determined. The number of tokens and the numbers of word forms, respectively, for each language are: Bulgarian 156,149 / 13,714, Old Church Slavonic 57,563 / 9,575, Polish 1498042 / 7,199, Slovak 10,6043 / 11,749, Slovenian 170,158 / 11,629, Ukrainian 122,275 / 9,789 and Latvian 208,965 / 17,046. We reduced aspect oppositions to the binary imperfective-perfective distinction, and took the difference of both occurrences. The differences were normalized, and ten thresholds between [.09:1] were set.

3 Results

We focused on the thresholds in the interval [.19, .59] in order to ensure a sufficient number of default and non-default encodings for the training of the SVM-classifier: the accuracy is almost independent of the threshold and thus of the frequency distribution: even with an almost equal distribution of default and nondefault aspect frequencies that is, with threshold .19, almost perfect accuracy values are achieved. The range of accuracy in [.19,.59] is: Bulgarian 99.5 – 99.8, Old Church Slavonic 94.3 – 97.8, Polish 99.7 – 99.9, Slovak 99.5 – 99.6, Slovenian 100 – 100, Ukrainian 99.1 – 100 and Latvian 98.3 – 99.5. Estimating UID-GLOBAL to our test set of languages, an identical pattern in all languages comes to light: the majority of variance values tends to be close to zero.

4 Conclusion

As Shannon's source coding theorem predicts, we found interactions of aspectual coding and information: Our study provides evidence that non-default coded verb forms are more informative than default forms. Almost identical accuracy has been achieved with all tested threshold values.

With regard to the second aim, our study discloses that UIDh holds within IC and SI: both features convey a uniform stream of information throughout the verb forms of the seven languages in focus.

The practical impact of our study concerns the assignment of word classes in languages such as Tagalog: default and non-default forms of a lemma correspond with different word classes.

References

Matthew Aylett and Alice Turk. 2004. The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous

Guiseppe Celano, Michael Richter, Rebecca Voll, and Gerhard Heyer. 2018. Aspect coding asymmetries of verbs: The case of Russian. *KONVENS 2018. PROCEEDINGS of the 14th Conference on Natural Language Processing*, 34 – 39.

Uriel Cohen Priva. 2008. Using information content to predict phone deletion. *Proceedings of the 27th West Coast Conference on Formal Linguistics*: 90 –98..

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5): 651 – 681.

Genzel, Dmitriy and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, pages 199 – 206.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*: 1 – 8.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61 (1): 23 – 62. doi: 10.1016/j.cogpsych.2010.02.002.

Thorsten, Joachims. 1998. *Text categorization with Support Vector Machines: Learning with many relevant features* Retrieved from http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS)*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition,* 106: 1126–1177.

Roger Levy. 2013. Memory and Surprisal in Human Sentence Comprehension. In Roger van Gompel, (ed.), *Sentence Processing*: 78 – 114. Psychology Press, Hove.

Steven T. Piantadosi, Harry Tily and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS*, 108(9): 3526 –3529.

Michael Richter, Yuki Kyogoku and Max Kölbl. 2019. Interaction of Information Content and Frequency as predictors of verbs' lengths. In Witold Abramowicz and Rafael Corchuelo (eds.), *Business Information Systems*. Springer, 271 – 282.

Claude E. Shannon, and Warren Weaver. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27.