Information from topic contexts: the prediction of aspectual coding of verbs in Russian

Michael Richter
Leipzig University
Institute of Computer Science
Augustusplatz 10, 04109 Leipzig
mprrichter@gmail.de

Tariq Yousef
Leipzig University
Institute of Computer Science
Augustusplatz 10, 04109 Leipzig
tariq@informatik.uni-leipzig.de

Abstract

The topic of this study is the prediction of aspectual coding asymmetries of verbs in Russian by the verbal feature *Average Information Content*. We employ the novel *Topic Context Model* that calculates the verbal information content from extra-sentential contexts i.e, the number of topics both in the target words' larger discourses and their local discourses. The former are the corpus, the latter are documents the target words occur in. In contrast to the study of Kölbl et al. (2020), TCM yielded disappointing results in this study. Our conclusion is that - compared to (Kölbl et al., 2020) - this is mainly due to the small number of local contexts we utilised.

1 Introduction

This paper reports the result of a study on the prediction of aspectual coding asymmetries in Russian, i.e., default and non-default aspectual coding of Russian verbs by the verb feature*Average Information Content* ($\overline{IC}$) (Priva, 2008), (Piantadosi et al., 2011). In general, the motivation for this study comes from Cysouw (Cysouw, 2005) who postulates that quantitative methods are needed for the examination of typological phenomena in languages. Our study focuses on the verbal coding asymmetry of perfective aspect and imperfective aspect, illustrated by a simple example from English: *eat up* normally carries perfective aspect, expressing that something is accomplished, like in *Max eats up the Lasagna*. However, in *Max is eating up the Lasagna*, the verb carries imperfective aspect expressing that the action is still ongoing. We calculated ($\overline{IC}$) in extra-sentential contexts of target words, emloying the novel *Topic Context Model* (TCM) (Kölbl et al., 2020).Evidence for a correspondence of aspectual coding of verbs and their information content comes from two previous studies on seven Slawic and Baltic languages (excluding Russian) (Richter and Yousef, 2019a,b). In contrast to the present study, the verbal information content in these studies was solely calculated on the basis of intra-sentential n-gram contexts. We argue, that 1-3-gram contexts are simply too small in order to model the cognitive activity of natural language processing. N-gram contexts of higher order cause serious problems because of sparseness of the occurrences of higher n-grams. In contrast, the point of departure of TCM is that the number of topics, both in a words overall and local discourses form a set of contexts of that word. The overall discourse is the collection of documents in a corpus. Local contexts can be the documents and texts of that corpus, in which the target word occurs in.

TCM outputs the average information content of a word w, given the topics of overall discourse and its local discourse. TCM is compatible with *surprisal theory* (Hale, 2001) (Levy, 2008) which requires large extra-sentential contexts for the calculation of information content. As baseline context models, we employ n-gram models, with context windows to either sides of the target words.

What kind of typological phenomenon is aspectual coding of verbs? Default aspectual coding of a verb marks the more frequent aspect type of a verb, in contrast, the non-default coding is the non-expected and thus more surprising aspect type (Richter and Yousef, 2019a), (Richter and Yousef, 2019b). Russian is a prototypical language for aspectual verb coding: Consider the verb nonebbl@id@@russianid@@russian 'write' in the UD corpus 'ru_syntagrus-ud-train' ('ud-treebanks-v2.1') (Nivre et al., 2017), that we exploited in this study. This verb has 225 occurrences in imperfective aspect and 138 occurrences in perfective aspect. The expectable aspect type is thus 'imperfective', we call it the *default* coded form, while the less probable aspect type is 'perfective', i.e., the non-default coded form. As regards morphology, non-default forms tend to be longer than the default forms (see for instance (?) and the Zipfian *principle of least effort* (Zipf, 1949) predicts that longer words - these tend to be the non-default forms - should be more informative than shorter ones.

2 Related work

The predictability of (physical) lengths of linguistic units by the feature 'information content' has been the subject in a number of studies: The interaction between phonetic duration and information has been disclosed for instance by utilising joint probability and conditional probability (Bybee and Scheibman, 1999; Gregory et al., 1999; Aylett and Turk, 2004; Pluymaekers et al., 2005). In a study on 10 Indo-European languages, Piantadosi et al. (Piantadosi et al., 2011) showed that average information content, calculated from n-gram contexts, is a strong predictor of of words lengths. Levshina (Levshina, 2017) proved for Arabic, Chinese, English, Finnish, German, Hindi, Persian, Russian and Spanish, that information which is calculated from the syntactic dependents of words is a good predictor of words lengths. The study of Celano et al. (Celano et al., 2018) on Russian confirm the results in Levchina (Levshina, 2017). In the study of Richter et al. (Richter et al., 2019) on 30 typologically diverse languages and Richter and Celano (Richter and Celano, 2019) on 18 diverse languages, information from n-gram contexts were better predictors of lengths of aspectual coded verbs than information from syntactic dependents of words.

3 Method

3.1 Corpus Description

As data resource, we used the *SynTagRus* corpus from UD[1]. This corpus is a compilation of a couple of corpora and consists of 522 documents and 48K sentences with an average length of 92 sentences per document. In order to model the topics in the corpus, we performed LDA and experimented with a various number of topics (50, 100, 200).

3.2 Default and non-default word forms and classifying

For the distinguishing of default from non-default forms, we adapted the method in (Richter and Yousef, 2019a). For each verb, the default and non-default aspect was determined. We normalised the differences and defined 10 thresholds between [.09:1] in order to define differences between default forms and non-default forms. A *Support Vector Machine* binary classifier has been employed in order to predict default and non-default aspectual verb forms (Joachims, 1998). We experimented with several features and decided to use average information content $\overline{IC}$, verb length (Celano et al., 2018) and the aspectual threshold as feature for the SVM classifier. We used 80% of the data set to train the model, and the rest to assess the quality of the classifier. The classification results based on information values by TCM were compared against the results based on $\overline{IC}$ by n-gram models which utilised quite as TCM the features length of verbs and thresholds (Richter et al., 2019) (Richter and Yousef, 2019a).

3.3 Latent Dirichlet Allocation and TCM

In order to detect the topics in the overall and the local contexts of the verbs, we employed *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). LDA is a statistical model in order to disclose the topics that appear in a text collection. In TCM, the contexts for the calculation of Information content of target words are defined as topics - disclosed by LDA. *Average Shannon Information content SI* for each target verb in the dataset is calculated, given the contexts / topics in which the verb occurs in, that is, given the topics both in the overall discourse and in its local discourses. The model workflow consists of three steps: (i) preprocessing to clean and prepare the dataset for the topic modelling step, (ii) LDA for topic detection (Blei et al., 2003) (iii) calculation of $\overline{IC}$.

Preprocessing:
First, the texts have been converted to lower case, subsequently they have been tokenised, and finally the stopwords and non-alphabetical tokens have been removed. With regard to our large data set and to the expectable long processing time, we applied neither stemming nor lemmatisation, since there is no evidence that this preprocessing steps could improve the results of topic modeling (Schofield and Mimno, 2016), (May et al., 2016)-

Latent Dirichlet allocation (LDA) (Blei et al., 2003):
The generative model LDA is based on the idea

[1] https://universaldependencies.org

that every document can be generated by a specific probability distribution of topics and that each topic is constituted by a specific probability distribution of the documents words. LDA aims to disclose patterns and contexts within a collection of documents collection and aims to classify documents by the feature ’topics’. The number of topics is a parameter of the LDA-algorithm and its output is a distribution of topics θ_i for each document d_i.

$\overline{IC}$ Calculation:
The calculation of the average information content $\overline{IC}$ for each target word has been carried out applying formula 1 where n is the number of contexts, (topics) of word w and $P(w|t_i)$ is the probability of word w given the context t_i.

$$\overline{IC}(w) = -\frac{1}{n}\sum_{i=1}^{n}\log_2 P(w|t_i) \tag{1}$$

4 Results

Table 1 displays the classification results of TCM models and the best performing n-gram model that is, a 3-gram model, with a context window of three words both to the left and to the right of a target word (’3L3R’).
For default forms, the n-gram model outperforms TCM and achieved the F1 score .98. TCM achieved lower F1 scores, that is, .83, .84 and .84 for 50, 100 and 150 topics, respectively. The same tendency came to light with non-default forms with, in general, lower F1 scores. The n-gram model achieved the F1 score .57 while for TCM, we observed the F1 scores .43, .40 and .32 for 50, 100 and 150 topics, respectively. Increasing the number of topics in TCM led to a decrease of F1 scores.
The study reveals considerable differences between the models respective the accuracy: TCM (50, 100 and 150 topics) reached an average accuracy of almost .75, the n-gram model reached accuracy of .96.

N-Gram model with 3L3R			
	Precision	Recall	F1
0	.97	.98	.98
1	.62	.53	.57
	Accuracy		.96
TCM with 50 Topics			
	Precision	Recall	F1
0	.78	.89	.83
1	.54	.35	.43
	Accuracy		.74
TCM with 100 Topics			
	Precision	Recall	F1
0	.79	.90	.84
1	.54	.32	.40
	Accuracy		.75
TCM with 200 Topics			
	Precision	Recall	F1
0	.76	.95	.84
1	.63	.21	.32
	Accuracy		.75

Table 1: Results of the SVM classifier based on the 3L3R-n-gram model and on TCM (50, 100 and 150 topics) predicting ’0’: default forms of Russian verbs and ’1’: non-default aspectual forms of Russian verbs.

5 Discussion and conclusion

The baseline n-gram model in this study provided moderate evidence for the correspondence of aspectual coding and both average information content and length of words. This finding corroborates the Zipfian principle of least effort. The n-gram based SVM-classifier achieved an almost maximum F1 score for default forms, achieved a higher F1 score for non-default forms than TCM and, additionally, a much higher accuracy. TCM on its turn performed poorly in the constellation of the Russian corpus. At first glance, TCM manages to classify correctly about three quarters of the verbs. However this percentage equals approximately the relation of default and non-default forms in the corpus. That is to say, the TCM based classifier did not manage to retrieve non-default forms. Our conclusion is that this is due to the small number of local contexts in the Russian corpus: TCM seems to work better with a high number of local discourses like in (Kölbl et al., 2020) because a large set of local discourses allow a better approximation to the ’true’ mean, i.e., the average information content. An additional difference in the data resources of the study in (Kölbl et al., 2020) and the present study was the thematic diversity in the Russian corpus as opposed to a greater thematic homogeneity in the IT technology-oriented Heise corpus.

In general, future research on TCM is desirable that focus (i) on the relevance of on the number of local contexts, and (ii) on the relevance of the thematic diversity within a corpus.

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 speech. *Language and speech*, 47(1):31–56.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Joan Bybee and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of don't in english. *Linguistics*, 37(4):575–596.

Giuseppe GA Celano, Michael Richter, Rebecca Voll, and Gerhard Heyer. 2018. Aspect coding asymmetries of verbs: The case of russian. In *KONVENS 2018. PROCEEDINGS of the 14th Conference on Natural Language Processing*, pages 34 – 39. Verlag der Österreichischen Akademie der Wissenschaften.

Michael Alexander Cysouw. 2005. Quantitative methods in typology. In *Quantitative Linguistik: ein internationales Handbuch= Quantitative linguistics*, pages 554–578. de Gruyter.

Michelle L Gregory, William D Raymond, Alan Bell, Eric Fosler-Lussier, and Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. In *Chicago Linguistic Society*, volume 35, pages 151–166.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Max Kölbl, Yuki Kyogolku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 2020. Keyword extraction in german: Information-theory vs. deep learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, Vol. 1*, pages 459 – 464.

Natalia Levshina. 2017. Communicative efficiency and syntactic predictability: A cross-linguistic study based on the universal dependencies corpora. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*, 135, pages 72–78. Linköping University Electronic Press.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Chandler May, Ryan Cotterell, and Benjamin Van Durme. 2016. An analysis of lemmatization on topic models of morphologically rich language. *arXiv preprint arXiv:1608.03995*.

Joakim Nivre, Lars Ahrenberg Zeljko Agic, et al. 2017. Universal dependencies 2.0 conll 2017 shared task development and test data. lindat/clarin digital library at the institute of formal and applied linguistics, charles university.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Mark Pluymaekers, Mirjam Ernestus, and R Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2-4):146–159.

Uriel Cohen Priva. 2008. Using information content to predict phone deletion. In *Proceedings of the 27th west coast conference on formal linguistics*, pages 90–98. Cascadilla Proceedings Project Somerville, MA.

Michael Richter and Giuseppe GA Celano. 2019. Aspectual coding asymmetries: Predicting aspectual verb lengths by the effects frequency and information content. *Topics in Linguistics*, 20(2):54–66.

Michael Richter, Yuki Kyogoku, and Max Kölbl. 2019. Interaction of information content and frequency as predictors of verbs' lengths. In *International Conference on Business Information Systems*, pages 271–282. Springer.

Michael Richter and Tariq Yousef. 2019a. Predicting default and non-default aspectual coding: Impact and density of information features. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Kaleidoscope Abstracts*, pages 275–277, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Michael Richter and Tariq Yousef. 2019b. Predicting default and non-default aspectual coding: Impact and density of information features. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA 2019), (extended version of 2019a)*.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.