Zero subject detection for Polish

Mateusz Kopeć
Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
m.kopec@ipipan.waw.pl

Abstract

This article reports on the first machine learning experiments on detection of null subjects in Polish. It emphasizes the role of zero subject detection as the part of mention detection – the initial step of end-to-end coreference resolution. Anaphora resolution is not studied in this article.

1 Introduction

Zero subject detection is an important issue for anaphora and coreference resolution for the null-subject languages, including all Balto-Slavic languages and most Romance languages. Their distinctive feature is the possibility for an independent clause to lack an explicit subject. Person, number, and/or gender agreement with the referent is indicated by the morphology of the verb:

(1) *Maria wróciła już z Francji. ØSpędziła tam miesiąc.*
"Maria came back from France. $\text{Ø}Had_{singular:feminine}$ spent a month there."

The recently created Polish Coreference Corpus[1] (PCC) (Ogrodniczuk et al., 2013) contains zero subject annotation. A markable representing the null subject is the verbal form following the position where the argument would have been expected. As tested on the development part of the corpus (described in detail later), omitting a personal pronoun is a frequent issue in the Polish language – about 30% of verbs do not have explicit subjects. Russo et al. (2012) reports similar figures for Italian (30.42%) and Spanish (41.17%).

Moreover, these null subjects are often part of large coreference clusters – the average size of a non-singleton coreference cluster in the development subcorpus was 3.56 mentions. At the same time, the non-singleton coreference cluster containing at least one zero subject had on average 5.89 mentions.

A mention detection module heavily influences the final coreference resolution score of an end-to-end coreference resolution system. In Ogrodniczuk and Kopeć (2011a) the system working on gold mentions achieved 82.90% F1 BLANC (Recasens and Hovy, 2011), whereas on system mentions the result dropped to 38.13% (the zero subject detection module was not implemented).

The aim of this paper is to find a method of automatic zero subject detection to improve the accuracy of mention detection as the initial step of coreference resolution.

2 Related Work

We present some of the most recent articles about machine learning zero subject detection.

Rello et al. (2012b) describes a Brazilian Portuguese corpus with 5665 finite verbs total, out of which 77% have an explicit subject, 21% a zero pronoun and 2% are impersonal constructions. They extract various verb, clause and neighboring token features for each verb occurrence and classify it into one of these 3 classes, achieving 83.04% accuracy of a decision tree learning classifier, better than the baseline result of the *Palavras* parser. A very similar study is conducted also for Spanish (Rello et al., 2012a), with the best result of the lazy learning classifier K^* (Cleary and Trigg, 1995) of 87.6% accuracy, outperforming the baseline of *Connexor* parser.

Chinese zero pronoun detection and resolution is presented by Zhao and Ng (2007). Features for zero pronoun identification consider mainly the gold standard parse tree structure. Their training corpus contained only 343 zero pronouns, as compared to 10098 verbs with explicit subjects – for Chinese, the phenomenon is much less frequent than for Polish or Spanish. Therefore they weigh

[1] Publicly available at http://zil.ipipan.waw.pl/PolishCoreferenceCorpus.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 221–225,
Gothenburg, Sweden, April 26-30 2014.

positive and negative examples to get the balance between precision and recall – the best result of 50.9% F_1 measure for positive to negative example weight ratio of 8:1 is reported.

A study for the Romanian language (Mihaila et al., 2011) describes a corpus consisting of 2741 sentences and 997 zero pronouns. Class imbalance is solved by training machine learning algorithms on all positive examples (zero pronouns) and the same number of negative examples (sampled from the corpus). Features used consider morphosyntactic information about the verb, precedence of the reflective pronoun "se" and the number of verbs in the sentence. Their best ensemble classifier scored 74.5% accuracy.

Only a few studies (for example (Broda et al., 2012; Ogrodniczuk and Kopeć, 2011b; Kopeć and Ogrodniczuk, 2012)) consider the problem of rule-based or machine learning coreference resolution for the Polish language, however these attempts leave zero subject detection as a non-trivial task for further study.

3 Problem statement

Table 1 presents part of speech definitions assumed in this article, based on the book about the National Corpus of Polish (Przepiórkowski et al., 2012). Coarse-grained POS indicates whether a word with a given part of speech may be a subject (*Noun*) or a verb (*Verb*) in a sentence. The last four columns present which morphosyntactic information is available for each part of speech. There are few differences in this definition with respect to the original approach in the book:

- We treat numerals, gerunds and pronouns as *Nouns* – because they are frequently subjects of the sentence and have the same morphosyntactic information as "standard" nouns.
- We do not consider *siebie* ("self", traditionally treated as pronoun) as a *Noun*, as it cannot be a subject.
- Tags: *impt, imps, inf, pcon, pant, pact, ppas, pred*, which are traditionally considered verb tags, are not treated by us as *Verbs*, because they cannot have a subject.

With such a definition of parts of speech, our task may be stated as follows: given a clause with a *Verb*, decide whether the clause contains a *Noun*

Coarse-grained POS	POS	Tag	Number	Case	Gender	Person
Noun	Noun	subst	+	+	+	
	Depreciative form	depr	+	+	+	
	Main numeral	num	+	+	+	
	Collective numeral	numcol	+	+	+	
	Gerund	ger	+	+	+	
	Personal pronoun – 1st, 2nd person	ppron12	+	+	+	+
	Personal pronoun – 3rd person	ppron3	+	+	+	+
Verb	Non-past form	fin	+			+
	Future *być*	bedzie	+			+
	Agglutinate *być*	aglt	+			+
	L-participle	praet	+		+	
	winien-like verb	winien	+		+	

Table 1: Parts of speech

which is the *Verb's* explicit subject. From now on in this paper, the words "noun" and "verb" have the meaning of *Noun* and *Verb*, respectively. In this study, we do not try to handle the cases of subjects not being nouns, as judging from our observations, it is very infrequent. We do take into account in our solution the cases of the subject not in the *nominative* case, as in the example:

(2) *Pieniędzy*$_{noun:genitive}$ *nie starczy dla wszystkich.*

"There wouldn't be enough money for everyone."

It is worth noting that Polish is a free-word-order language, therefore there are many possible places for the subject to appear, with respect to the position of the verb.

As the corpus has only automatic morphosyntactic information available (provided by the PANTERA tagger (Acedański, 2010)), not corrected by the coreference annotators, the only verbs considered in this study are the ones found by the tagger. If such a verb was marked as a mention by the coreference annotator (*verb mention* in table 2), it is a positive example for our machine learning study, otherwise a negative one. Sentence and clause segmentation in the corpus was also automatic. We are aware that the corpus used for the study was not perfectly suited for the task – verbs with a zero subject are not marked there explicitly, but can only be found based on automatic tagging. However the tagging error of detecting verbs is reported as not higher than 0.04% (for the *fin* tag, see (Acedański, 2010) for details), so we consider the resource sufficiently correct.

4 Development and evaluation data

Each text of the Polish Coreference Corpus is a 250-350 word sample, consisting of full, subsequent paragraphs extracted from a larger text. Text genres balance correspond to the National Corpus

Corpus	# texts	# sentences	# tokens	# verbs	# mentions	# verb mentions
Development	390	6481	110379	10801	37250	3104
Evaluation	389	6737	110474	11000	37167	3106
Total	779	13218	220853	21801	74417	6210

Table 2: Zero subject study data statistics

of Polish (Przepiórkowski et al., 2012). At the time this study started, 779 out of 1773 texts (randomly chosen) of the Polish Coreference Corpus were already manually annotated. Annotated texts were randomly split into two equal-sized subcorpora for development and evaluation. Their detailed statistics are presented in Table 2.

4.1 Inter-annotator agreement

210 texts of the Polish Coreference Corpus were annotated independently by two annotators. This part was analyzed for the inter-annotator agreement of deciding if a verb has a zero subject or not. In the data there were 5879 verbs total, for which observed agreement yielded 92.57%. Agreement expected by chance (assuming a per annotator chance annotation probability distribution) equalled 57.52%, therefore chance-corrected Cohen's κ for the task equalled 82.51%.

4.2 Results of full dependency parsing

The first Polish dependency parser was recently developed and described by Wróblewska (2012). The author reports 71% LAS[2] and 75.2% UAS[3] performance of this parser. This parser was used to detect null subjects – every verb lacking the dependency relation of the subject type (`subj`) was marked as missing the subject. This baseline method achieved accuracy of 67.23%, precision of 46.53%, recall of 90.47% and F_1 equal to 61.45%. These results are worse than a simple majority baseline classifier, therefore current state-of-the-art Polish dependency parsing is not a satisfactory solution to the task stated in this article.

5 Features

Based on a number of experiments on the development corpus, we chose a number of features presented in table 3.

Subject candidate existence features from the bottom of the table 3 use variables: c_1, c_2 and w. Separate feature was generated for each combination of these three variables. The variable w represents the window around the verb, with following values: the clause containing the verb, the sentence containing the verb, windows of 1 to 5 tokens before the verb, windows of 1 to 5 tokens after the verb, windows of 1 to 5 tokens both before and after the verb. Variable c_1 represents compatibility of noun and verb, with values being any nonempty subset of the set of following conditions: case of the noun equal to *nominative* (NOM), number agreement with the verb (NUM), person or gender agreement (POG), depending on which was available to check, see Table 1. Variable c_2 is similar to c_1, with the following values: {NOM}, {POG}, {NOM, POG}.

Feature	Type
Verb features	
number of the verb – to help with cases of plural verbs having two or more singular nouns as subject	nominal
tag of the verb – as it may happen, that some parts of speech behave differently	boolean
is the verb on the pseudo-verbs list extracted from (Świdziński, 1994) – i.e. may not require a subject	boolean
Neighboring token features	
tag of the next token	nominal
tag of the previous token	nominal
is the previous tag equal to *praet* – a redundant feature to the previous one, but it should help with the cases like: $\ldots$ *była*$_{praet}$*m*$_{aglt:pri}$ $\ldots$ "$\ldots$ *(I) was* $\ldots$" when we split a word into a L-participle and agglutinate. Annotation guidelines were to only mark the agglutinate as a mention, when the verb does not have an explicit subject	boolean
does one of the previous two tokens have the *pred* tag – should allow detecting examples similar to: *Można*$_{pred}$ *się było*$_{praet}$ *tego spodziewać.* "$\ldots$ *It could have been expected.* $\ldots$" *Trzeba*$_{pred}$ *było*$_{praet}$ *myśleć wcześniej.* "*(One) should have thought before.*" when *było ("have")* cannot have subject, as it is part of an impersonal construction	boolean
is the next tag *inf* – similar role to the previous feature, as in: *Wtedy należy*$_{fin}$ *poprosić*$_{inf}$. "*(One) should then ask for it.*" when *należy ("one should")* cannot have a subject	boolean
is the previous token a comma	boolean
Length features	
number of tokens in the sentence (following the hypothesis, that the shorter the sentence/clause, the less likely for the subject to appear)	numerical
number of tokens in the clause with the verb	numerical
Subject candidate existence features	
existence of a noun not preceded by *jak/jako ("as")* in window w fulfilling conditions from set c_1	boolean
existence of at least two nouns not preceded by *jak/jako ("as")* in window w both fulfilling conditions from set c_2	boolean

Table 3: Features

6 Evaluation

Presented features were used to train a machine learning algorithm. We chose the JRip implementation of RIPPER (Cohen, 1995) from WEKA (Hall et al., 2009) for the possibility to interpret the rules, which is outside of the scope of this paper.

6.1 Accuracy on the development corpus

A baseline model which always predicts that a verb has an explicit subject achieves 71.13% ac-

[2]Labeled attachment score – the percentage of tokens that are assigned a correct head and a correct dependency type.

[3]Unlabeled attachment score – the percentage of tokens that are assigned a correct head.

		True values	
		null subject	explicit subject
Predictions	null subject	2093	815
	explicit subject	1013	7079

Table 4: Confusion matrix

curacy on the development data. The upper bound of the ITA (as stated earlier) is around 92.57% accuracy.

We used 10-fold cross-validation which was repeated 10 times with different random seeds for training and train/test splits. The average from the total of 100 trials (each cross-validation split separately) was equal to 82.74%, with standard deviation of 1.27%. As the Shapiro-Wilk (1965) test for normality for this data gives p-value of 0.38, it may be assumed that it follows the normal distribution. In that case, the 95% confidence interval for the accuracy is equal to [82.49%, 82.99%].

6.2 Accuracy on the evaluation corpus

The evaluation corpus was used only for two experiments presented below: to calculate accuracy and learning curve of the developed solution.

We used the model learnt on the development corpus and tested it on the evaluation corpus, achieving 83.38% accuracy. A majority classifier would achieve 71.76% accuracy on this corpus. The confusion matrix is depicted in Table 4. For finding the null subjects, recall of 67.39% and precision of 71.97% gives F_1 measure of 69.60%.

6.3 Learning curve

To test how the number of training examples influences the quality of the trained classifier, we used subsets of the development corpus of various sizes as training sets. The test set was the same in all cases (the evaluation corpus). Proportions of the examples used ranged from 5% to 100% of the development corpus, each proportion was tested 10 times to provide an estimation of variance. For example, to evaluate the efficiency of the classifier trained on 5% of the training examples, we randomly sampled 5% of the examples, trained the classifier and tested it on the full evaluation corpus. Then we repeated it another 9 times, randomly choosing a different 5% portion of the examples for training.

Again the Shapiro-Wilk test was taken to assess the normality of results for each proportion, out of 19 proportions tested (the proportion of 1 was of course not tested for normality), only 3 had p-value less than 0.1, therefore we assumed that the data is distributed approximately normally. The 95% confidence intervals of the classifiers trained on a given proportion of the development corpus are shown in the Figure 1. The algorithm clearly benefits from having more training examples. We observe that the curve is generally of the desired shape, yet it flattens when approaching the full training set used. It may suggest that the developed solution would not be able to significantly exceed 84%, even given more training examples.

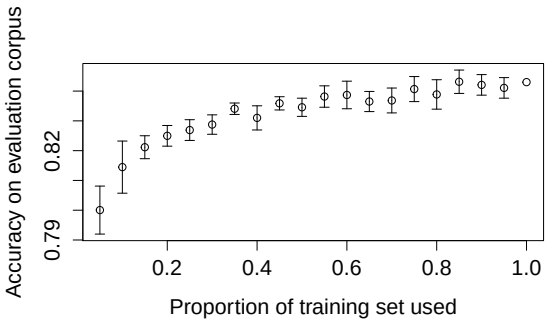

Figure 1: Learning curve

7 Conclusions and future work

This article presented an efficient zero subject detection module for Polish. We highlighted some difficult examples to take into account and proposed a solution for the Polish language.

The achieved accuracy of 83.38% significantly exceeds the baseline of majority tagging, equal to 71.76%, but there is still room for improvement, as the upper bound of 92.57% was computed. The achieved result for the task of null subject detection looks promising for the application in mention detection for coreference resolution.

The invented solution needs to be incorporated in a complete coreference resolver for Polish and evaluated for the extent to which using such an advanced separate classifier for zero subject detection improves the mention detection and, furthermore, end-to-end coreference resolution accuracy.

Acknowledgements

The work reported here was cofounded by the Computer-based methods for coreference resolution in Polish texts project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40) and by the European Union from resources of the European Social Fund. Project PO KL „Information technologies: Research and their interdisciplinary applications”.

References

Szymon Acedański. 2010. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Bartosz Broda, Łukasz Burdka, and Marek Maziarz. 2012. IKAR: An Improved Kit for Anaphora Resolution for Polish. In *COLING (Demos)*, pages 25–32.

John G. Cleary and Leonard E. Trigg. 1995. K*: An instance-based learner using an entropic distance measure. In *In Proceedings of the 12th International Conference on Machine Learning*, pages 108–114. Morgan Kaufmann.

William W. Cohen. 1995. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

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

Mateusz Kopeć and Maciej Ogrodniczuk. 2012. Creating a Coreference Resolution System for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 192–195, Istanbul, Turkey. ELRA.

Claudiu Mihaila, Iustina Ilisei, and Diana Inkpen. 2011. Zero Pronominal Anaphora Resolution for the Romanian Language. *Research Journal on Computer Science and Computer Engineering with Applications" POLIBITS*, 42.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011a. End-to-end coreference resolution baseline system for Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 167–171, Poznań, Poland.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011b. Rule-based coreference resolution module for Polish. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 191–200, Faro, Portugal.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2013. Polish coreference corpus. pages 494–498.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw.

Marta Recasens and E. Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. pages 485–510.

Luz Rello, Ricardo Baeza-Yates, and Ruslan Mitkov. 2012a. Elliphant: Improved Automatic Detection of Zero Subjects and Impersonal Constructions in Spanish. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 706–715, Avignon, France, April. Association for Computational Linguistics.

Luz Rello, Gabriela Ferraro, and Iria Gayo. 2012b. A First Approach to the Automatic Detection of Zero Subjects and Impersonal Constructions in Portuguese. *Procesamiento del Lenguaje Natural*, 49:163–170.

Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. 2012. Improving machine translation of null subjects in Italian and Spanish. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–89, Avignon, France, April. Association for Computational Linguistics.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, Dec.

Marek Świdziński. 1994. Syntactic dictionary of polish verbs.

Alina Wróblewska. 2012. Polish dependency bank. *Linguistic Issues in Language Technology*, 7(1).

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *EMNLP-CoNLL*, pages 541–550. ACL.