Automatic Detection of Annotation Errors in Polish-language Corpora

Łukasz Kobyliński

Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland,
lkobylinski@ipipan.waw.pl

Abstract. In this article we propose an extension to the variation n-gram based method of detecting annotation errors. We also show an approach to finding anomalies in the morphosyntactic annotation layer by using association rule discovery. As no research has previously been done in the field of morphosyntactic annotation error correction for Polish, we provide novel results based on experiments on the largest available Polish language corpus, the National Corpus of Polish (NCP). We also discuss the differences in the approaches used earlier for English language data and the method proposed in this article, taking into account the characteristics of Polish language.

1 Introduction

Annotated text corpora are one of the most important resources used in linguistics. Particularly, in computational linguistics, they serve as a basis for training automated taggers, as well as may be used as a source of information for speech recognition and machine translation systems. These corpora are either annotated manually by qualified linguists, or automatically, using taggers. Unfortunately even the most recent automated taggers are far from being 100% accurate. For example, in the case of part-of-speech tagging of Polish texts, the best-performing automated taggers achieve "weak correctness" (measured as the percent of words for which the sets of interpretations determined by the tagger and the gold standard are not disjoint) of between 91.06% (TaKIPI, [1]) to 92.44% (PANTERA, [2]). The reliability of annotation has a direct impact on the results of most other language-related research, as the methods used there usually rely on corpora and their annotation to perform their tasks. There is thus a burning need to improve the tagging accuracy, as each incorrectly annotated word potentially lowers the results of other, higher-level text processing techniques.

As manual correction of errors in the entire corpus is impractical, it is therefore necessary to employ an automated method of tagging error detection in the corpus to filter only potential mistakes and present them to human annotators.

2 Previous Work

Below we discuss some of the prominent representatives of the approaches previously proposed to the problem of annotation error detection.

Dickinson and Meurers [3] show an effective approach to annotation error detection, which is based on the idea of finding "variation n-grams" in a corpus. Variation n-grams are sequences of tokens, which appear multiple times in the text and contain at least one word that has been assigned different annotation tags throughout the corpus. The word with ambiguous annotation is called the "variation nucleus" and is a potential place, where an annotation error might have occurred. The n-grams are discovered in the corpus using an incremental approach: at first unigrams are found and their position stored; next, each unigram is extended left or right by one word (if possible) and the resulting n-gram stored; the second step is repeated until no n-gram can be further extended. It is thus a method of finding the largest contexts of words, for which a tagging error might have been introduced during the annotation process.

A method of using association rules directly mined from the corpus data to find frequent relationships between the annotations of segments appearing in similar contexts has been proposed by [4]. Rules with high confidence and support have then been used to detect word occurrences, which violate these strong rules. The authors have concluded that the method achieved ca. 20% precision and that the limiting factor was the sparse annotation of the Czech PDT corpus.

There are no published works dealing with automated detection of morphosyntactic annotation errors in Polish language corpora. Having in mind the fact that very large corpus of Polish has just been released (the National Corpus of Polish [5] is a reference corpus of Polish language containing over fifteen hundred millions of words) and that they are used regularly in most other projects related to language processing, it is an important research problem to provide such a method, which would provide accurate results for Polish and to improve the approaches already described in the literature for other languages.

3 Variation N-grams in Annotation Error Detection

A variation n-gram ([3]) is an n-gram of words from a collection of texts, which contains one or more words annotated differently in another occurrence of the n-gram in the corpus. For example, the following is a variation 9-gram taken from the manually annotated 1 million word subcorpus of the NCP:

- Zamykam dyskusję. Do **głosowania** [głosować:ger:sg:n:imperf:n] nad uchwałą Senatu przystąpimy *jutro rano.*
  I close the discussion. We will proceed to **a vote** on the resolution of the Senate *tomorrow morning.*
- Zamykam dyskusję. Do **głosowania** [głosowanie:subst:sg:gen:n] nad uchwałą Senatu przystąpimy *w bloku głosowań.*
  I close the discussion. We will proceed to **a vote** on the resolution of the Senate *in a series of votings.*

The word "głosowania" in above example is annotated as a gerund in one occurrence, while in another occurrence it is tagged as a noun. Dickinson and

Meurers call such a word a "variation nucleus", as it constitutes a variation n-gram, which indicates an existence of inconsistency in corpus annotation.

In the original formulation of the algorithm, a variation n-gram is created by first finding words in a corpus, which have exactly the same orthographic form, but different annotation. Such unigrams are then extended to neighboring words, if their orthographic form appears in more than one fragment. Application of this method to the manually annotated 1 million subcorpus of the NCP resulted in finding variation n-grams of length up to 67. Intuitively, longer variation n-grams, representing more similar contexts with annotation anomaly, are the most promising candidates for annotation errors. Unfortunately, the vast majority of unique n-grams found was not longer than 6 words. Unique n-grams are understood as n-grams, which are not contained by any longer n-gram discovered.

To evaluate the actual accuracy of the method, we have firstly prepared a list of annotation errors spotted and corrected manually in the corpus by a trained linguist. In course of his work the linguist corrected 2 692 mistakes in the corpus, of which 1 332 corrections considered the morphosyntactic annotation layer. We have used this information to estimate the recall of the approach, understood as the fraction of previously found annotation mistakes in the corpus, which were also detected by the automatic method. Table 1 presents the number of manually corrected segments, which have also been detected by the variation n-gram approach.

Table 1. Errors detected automatically vs errors corrected manually in the corpus; minN – minimum length of variation n-grams that were inspected, TP – true positives (among the 1 332 manual corrections), FP – false positives, F – value of the F measure.

minN	suspicious segments	TP	FP	precision	recall	F
3	54970	398	38	0.72%	29.88%	1.41%
4	10448	97	3	0.93%	7.28%	1.65%
5	2513	24	0	0.96%	1.80%	1.25%
6	873	12	0	1.37%	0.90%	1.09%

We have also performed a direct evaluation of the precision of the method, by inspecting manually the list of possible annotation mistakes produced by the algorithm. The results of such an experiment are presented in Table 2.

The previously stated intuition that longer n-grams have a much greater probability of indicating an actual annotation error is clearly backed by the experimental data, as precision of variation n-grams longer than 10 surpasses 70%, while global average was 52.55%. Another intuition, suggested by the authors of [3], is that variation nuclei appearing on a verge of a variation n-gram are usually not an annotation error, as the context is different on that side of the n-gram. We have repeated such an experiment, including only non-verge variation n-grams and the results show an increase in precision of the method, but at a cost of lower recall (see Table 3).

Table 2. Manual verification of the list of errors detected automatically; N - length of variation n-grams, verified – number of manually verified contexts, errors – number of actual annotation errors.

N	suspicious contexts	verified	errors	precision
4	1192	19	10	52.63%
5	373	9	5	55.56%
6	104	21	9	42.86%
7	32	16	11	68.75%
8	24	15	5	33.33%
9	23	20	6	30.00%
>=10	37	37	26	70.27%
	1785	137	72	52.55%

Table 3. Errors detected automatically using the non-fringe heuristic vs errors corrected manually in the corpus.

minN	segments	TP	FP	precision	recall	F
3	18855	203	10	1.08%	15.24%	2.01%
4	4870	73	2	1.50%	5.48%	2.35%
5	1605	23	0	1.43%	1.73%	1.57%
6	678	11	0	1.62%	0.83%	1.09%

4 Increasing Recall of the N-gram Detector

Experiments with the original annotation error detection method proposed by [3] have shown a difficulty in the direct application of the approach to Polish language texts. The number of discovered variation n-grams in corpora of similar sizes is much lower for Polish than it is for English. As Polish is inflectional, the number of n-grams that can be built on the basis of orthographic word forms is far more limited than for English. It thus possible to achieve similar precision ratio as for English, but the number of detected suspicious contexts and consequently the (estimated) recall is much lower. Based on the original variation n-gram method, here we propose modifications to increase the recall of the approach and make the algorithm more suitable for inflectional languages.

Firstly, we have experimented with generalization of certain word types, by eliminating the need of two words to have exactly the same orthographic form to be included in an n-gram. For example, in case of punctuation, abbreviations and numbers the exact word form used should not differentiate two similar contexts. Table 4 shows the results of experiments, in which n-grams have been extended to neighboring words of such types, regardless of their orthographic form (e.g. an n-gram has been extended to include a comma, even if in another context a period was used in that place).

Secondly, we have experimented with building variation n-grams based solely on the part-of-speech tags of words, ignoring their orthographic form. In such a scenario we assume that similar sequences of POS tags represent contexts, having similar grammatical structure. Table 5 presents the results of error detection

Table 4. Errors detected automatically vs errors corrected manually in the corpus; orthographic form of *interp*, *brev*, *num*, *numcol* types ignored.

minN	segments	TP	FP	precision	recall	F
4	8939	90	2	1.01%	6.76%	1.75%
5	2878	32	0	1.11%	2.40%	1.52%
6	1107	16	0	1.45%	1.20%	1.31%

using that approach. Clearly, the recall of the method has successfully been increased, at a cost of lower precision.

Table 5. Errors detected automatically vs errors corrected manually in the corpus; n-grams extended based on their POS tags.

minN	segments	TP	FP	precision	recall	F
4	28499	257	30	0.90%	19.29%	1.72%
5	9547	98	9	1.03%	7.36%	1.80%
6	2762	36	0	1.30%	2.70%	1.76%

5 Detecting Anomalies in Annotation using Association Rules

Association rule mining has been proposed in [6], originally as a method for market basket analysis. This knowledge representation method focuses on showing frequent co-occurrences of attribute values in data. Based on the original idea of [4], we have used association rule mining to identify relationships in corpus morphosyntactic annotation, which were of very high confidence, but still not equal to 100%. This allowed us to detect word-annotation pairs, which were suspiciously rare and therefore could constitute an error. We have mined rules having support greater or equal to 0.1% and confidence above 99% in a random sample of corpus contexts. We have then transformed the discovered rules into search queries, allowing us to identify instances, which did not support the 99% confident rules. Given a rule of the form:

$$attr_1, \ldots, attr_n \longrightarrow attr_{n+1}, \ldots, attr_m,$$

we have formed a search query as follows:

$$attr_1 \ \& \ \ldots \ \& \ attr_n \ \& \ !attr_{n+1} \ \& \ \ldots \ \& \ !attr_m.$$

Below we give an example of several rules mined from the corpus and associated search queries, along with the number of actual errors identified by the query. Numbers in parenthesis indicate the number of segments that supported the rule antecedent / rule consequent.

- base=my ⟶ ctag=ppron12 (276/274)
  - query [base=my&pos!=ppron12] returns 1 error in 7 results,
- ctag=aglt ⟶ base=być (290/288)
  - query [pos=aglt&base!=być] returns 10 errors in 24 results,
- base=no msd=[null] ⟶ ctag=qub (446/442)
  - query [base=no&pos!=qub] returns 2 errors in 13 results,
- base=tak ctag=adv ⟶ msd=pos (118/117)
  - query [base=tak&pos=adv&degree!=pos] returns 27 errors in 27 results.

6 Conclusions and Future Work

We have presented experimental results of two approaches to automatic detection of annotation errors applied to the National Corpus of Polish, a reference linguistic resource for Polish. We have successfully adapted methods proposed earlier for English language corpora to inflectional Polish language and proposed extensions, which may be used to increase recall of the detector, regardless of the target language. Described approaches to automatic detection of annotation errors proved to reduce the amount of time needed to identify mistakes and facilitated correction of a large corpus, namely the National Corpus of Polish.

In the future, we plan to combine various detection methods to further improve both the precision and recall of the system. As each of the approaches may identify different contexts as potentially erroneous, aggregating their results is a promising direction of further work.

7 Acknowledgements

The author would like to thank Łukasz Szałkiewicz for his linguistic work and Michał Lenart for sharing algorithm implementations. The work has been funded by the National Science Centre project number DEC-2011/01/N/ST6/01107.

References

1. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly **11**(1–2) (2007) 151–167
2. Acedański, S.: A morphosyntactic Brill tagger for inflectional languages. In: Advances in Natural Language Processing. (2010) 3–14
3. Dickinson, M., Meurers, D.: Detecting errors in part-of-speech annotation. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), Budapest, Hungary (2003)
4. Novák, V., Razímová, M.: Unsupervised detection of annotation inconsistencies using apriori algorithm. In: Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore, Association for Computational Linguistics (August 2009) 138–141
5. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., eds.: Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw (2012)
6. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C., USA (May 1993) 207–216