The Polish Summaries Corpus

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

Abstract

This article presents the Polish Summaries Corpus, a new resource created to support the development and evaluation of the tools for automated single-document summarization of Polish. The Corpus contains a large number of manual summaries of news articles, with many independently created summaries for a single text. Such approach is supposed to overcome the annotator bias, which is often described as a problem during the evaluation of the summarization algorithms against a single gold standard. There are several summarizers developed specifically for Polish language, but their in-depth evaluation and comparison was impossible without a large, manually created corpus. We present in detail the process of text selection, annotation process and the contents of the corpus, which includes both abstract free-word summaries, as well as extraction-based summaries created by selecting text spans from the original document. Finally, we describe how that resource could be used not only for the evaluation of the existing summarization tools, but also for studies on the human summarization process in Polish language.

Keywords: corpus, summarization, Polish, annotation

1. Introduction

The attempts of automatically creating summaries of Polish texts is not new; it has been theoretically discussed since 2000 e.g. by (Branny and Gajęcki, 2005; Dudczak et al., 2008a; Dudczak et al., 2008b; Dudczak et al., 2010; Głowińska and Głowiński, 2003) and resulted in implementation of several extractive tools:

- *PolSumm* (Ciura et al., 2004; Suszczańska and Kulików, 2003),
- *Lakon* (Dudczak, 2007), see also `http://www.cs.put.poznan.pl/dweiss/research/lakon/index_en.html`,
- *Świetlicka's Summarizer* (Świetlicka, 2010), see also `http://clip.ipipan.waw.pl/Summarizer`,
- *OpenTextSummarizer* (Rotem, 2003), used as a baseline summarizer by the Applied Technology for Language-Aided CMS project (ATLAS) language processing platform (Ogrodniczuk and Karagiozov, 2011),
- discourse-centered multilingual summarizer (Anechitei et al., 2013) implemented as a target solution for ATLAS.

Three of them (*Świetlicka's Summarizer*, *Lakon* and *OpenTextSummarizer*) are currently available in *Multiservice*, a demonstration platform for the Polish language tools (Ogrodniczuk and Lenart, 2013). At the same time until now there existed no data which could be used to carry out a formal evaluation of the summarizers and provide a quality comparison of the rival tools.

This paper intends to present such a new resource created to support the development and evaluation of the tools for automated summarization of Polish — the Polish Summaries Corpus (see `http://zil.ipipan.waw.pl/PolishSummariesCorpus`) — and encourage its use for the evaluation of summarization tools.

2. Related work

Of course, the summarization tools available for Polish were tested by their authors, and some of them created test corpora for the task. *Lakon* summarization tool was evaluated with a corpus of sentence-extraction-based summaries of the 10 newspaper texts, summarized by as many as 30 independent annotators for a single document. There was only one summary size, 20% of the original sentence count. This resource – although very interesting for the comparison of the inter-annotator agreement – is too small for the general evaluation and limits the possibility of testing summarization systems to the ones based on the sentence selection technique.

Similar limitations apply to the corpus created by Świetlicka (2010): it contained larger number of press articles summarized (169), but each one had only a single summary. Such summary consisted of a selection of 30% of the most informative sentences, and half of them was marked as the most informative from this subset. Again, with that corpus we may only evaluate and compare the sentence extraction-based algorithms.

3. Corpus desiderata

As an conclusion from the analysis of the existing corpora of Polish summaries, we have designed the following desiderata for the new corpus:

- it should contain as large number of texts as possible, not fewer than a few hundred,
- there should be many sizes of summaries for each text, to allow for testing the behaviour of algorithms in different compression settings,
- it should contain extractive summaries, but not limited to sentence selection, but rather word selection, to allow for research on human summarization techniques,

- it should contain also abstractive summaries, written without any constraints imposed on the annotators, to be able to test the evaluation measures in such setting and also study the human summarization process,
- each summary should have many versions, written by different, independent annotators, to overcome a single-annotator bias.

The corpus presented in this article fulfils these desiderata.

4. Corpus Source and Preprocessing

The Polish Summaries Corpus contains manual single-document summaries of press articles. This section presents the procedure for obtaining the texts, which were manually summarized.

4.1. The Original Data

Texts of the corpus were derived from the "Rzeczpospolita corpus" (RC) (Presspublica, 2002) — a collection of articles from the Web archive of Rzeczpospolita, a nationwide Polish daily newspaper. RC consists of 190 379 pseudo-HTML files (1.9 GB data) dating from 1993 to 2002, with unequal representation of individual years. The data set has been made available by its owners (Presspublica, the publisher of the newspaper) for research and so far they have been used many times in various computational linguistic tasks.

Every file in RC contains one or more articles (or practically none, when it references some non-textual content, such as a comic strip). Textual data is accompanied by HTML metadata (not always complete), such as the name of the newspaper section (`DZIAL`) in which the article was published, e.g.:

```
<META NAME="DZIAL"
      CONTENT="gazeta-sport">
```

where `gazeta-sport` can be translated as 'newspaper-sport'. This section information, whenever filled in (empty for 8 165 files) was used to detect text domains (106 variants).

4.2. Data Selection and Conversion

Since the HTML code of the files in RC is not valid (particularly it does not contain an `<html>` tag), article borders have been detected using simple heuristics based on the verified assumption that particular HTML comments, output by the Presspublica archiving system, mark the beginning and end of each text. Aggregate and 'empty' texts have been removed from our result set by counting HTML elements representing document title (`<FONT SIZE="5">...</FONT>`). For the sake of our experiment, all texts have been finally converted to plain text and certain HTML content was completely removed (such as `<TABLE>...</TABLE>` or `<MENU>...</MENU>`).

By limiting the resulting data set to domains represented by more than 1000 articles sized between 1000 and 4000 words, 7 most frequent domains were selected. Number of selected domains was chosen to have at least 30 texts in each one. In the last step of the data selection the articles were manually investigated to remove aggregates, interviews, legal acts or sports results (frequently published in the form of articles, but not suitable for the typical single-document news article summarization task).

Text domain	**Abstractive corpus**	**Extractive corpus**
Social and political	22	393
Sport	22	36
Economy	22	34
Cultural news	22	32
Law	22	26
National news	22	24
Science and technology	22	24
Total	**154**	**569**

Table 1: Selected domains

Out of these texts 569 were manually summarized: all of them have the extractive summaries, 154 out of 569 have also the abstractive summaries. The details about the summarization process are presented in the next section. Table 1 gives an insight into the distribution of the selected domains among the summarized texts. Because all the texts with abstractive summaries have also extractive summaries, one may use the corpus of 154 texts if he needs summaries of both kinds, while if only extractive summaries are required, one may benefit from larger, 569-text corpus.

Number of texts annotated was limited to 569 because of time and cost constraints, and the larger number of extractive summaries is due to the fact, that most of the automatic summarization systems are based on extraction techniques. Majority of texts in the extractive corpus is from social and political domain, as the number of texts from other domains in the "Rzeczpospolita corpus" was not enough to maintain the equal ratio of each type, as in the abstractive corpus.

5. Manual Summarization

Manual summarization was conducted by 11 annotators, which were randomly assigned texts to summarize. They were using three dedicated applications: for acquiring texts to work on (available at `http://zil.ipipan.waw.pl/DistSys`), for creating abstractive summaries, and for creating extractive summaries (both available at `http://zil.ipipan.waw.pl/SummaryAnnotationTools`).

Extractive summary annotation tool is depicted in Fig. 1. It shows both the original text and the summary and facilitate selection of fragments as well as counting percentages on the fly. Three tabs allow for annotation of three summaries of different sizes for the text loaded. Similar application was used for the abstractive summary annotation.

5.1. Extractive summaries

Annotators were instructed to create three extractive summaries of a given text, each constituting approximately 20%, 10% and 5% of the word count of the original (for a 1000-word source text the resulting summaries should

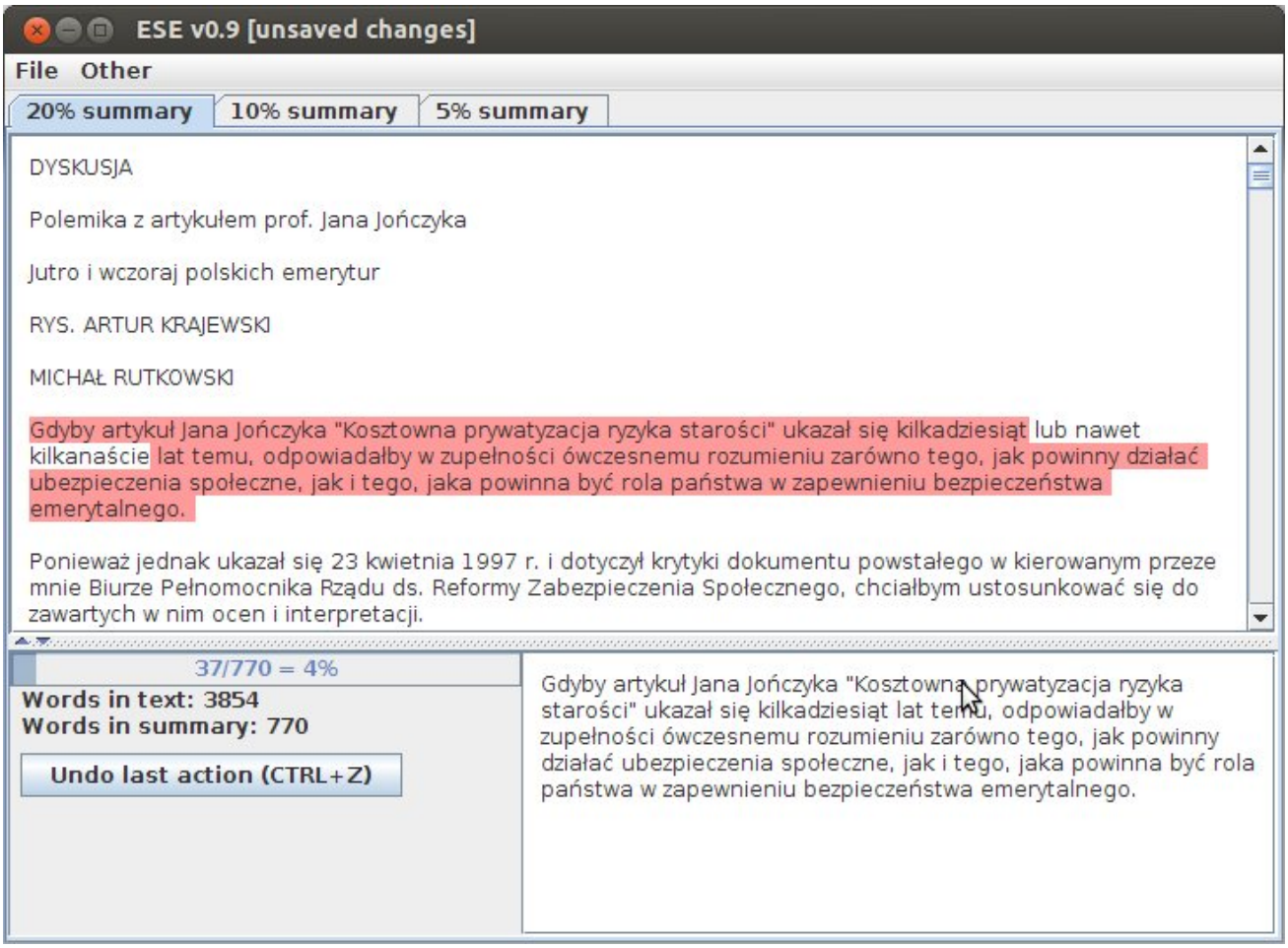

Figure 1: Application for extractive summaries annotation

then respectively be 200, 100 and 50 words). Minor (a few word-length) deviations were acceptable to encourage annotators to select the most important fragments — and not the ones which would add up to the desired limit.

Only original words and punctuation in the original order had to be used (so that annotators could e.g. select just the superordinate clause and a finishing dot, removing the less important part of a sentence such as subordinate clauses, interjections, excessing adjectives — but not creating abbreviations from first letters of a proper name MWU). No document title, subtitle or author should be included, neither any information referring to the summarization process (such as "the text explains..."). The resulting summary was supposed to be grammatically correct and coherent, but tricks such as linking two phrases from two sentences with a conjunction coming from a third one were discouraged. As phrases could be selected and sentences combined, lowercase start of the sentence or an uppercase character in the middle of the resulting sentence was acceptable.

The sequence of summaries was supposed to be inclusive, i.e. the 10-percent summary had to use only fragments previously selected for a 20-percent summary — and, similarly, the 5-percent summary had to use only fragments previously selected for a 10-percent summary. In this way a partial ranking of sentences could be inferred.

5.2. Abstractive summaries

Similarly to the previous task, annotators were instructed to create 3 abstractive summaries of a given text, each constituting approximately 20%, 10% and 5% of the word count of the original, with acceptable minor deviations in word count.
Contrary to extractive summaries, abstractive summaries did not have to contain fragments of original texts and could express the same ideas "in own words" of an annotator. Similarly, longer summaries could (but did not have to) contain fragments of shorter ones.

5.3. Independent annotations

Based on the opinions of many researchers, that there is no single "gold" summary for a given text (see for example one of the seminal works in the domain – (Rath et al., 1961)), we decided to provide 5 independent versions of the summaries described above, each one written by a different annotator (yet single annotator always summarized to reach all three sizes: 5%, 10% and 15%). We have chosen 5 versions following the research of Nenkova, where 4 to 5 summaries is said to provide an optimal balance of annotation effort and reliability for the Pyramid method evaluation (see for example (Nenkova et al., 2007)).
Therefore, because of 3 summary sizes for each text (20, 10 and 5%), our corpus contains altogether $569 * 3 * 5 = 8535$

extractive summaries and $154 * 3 * 5 = 2310$ abstractive ones, with makes a total of 10845 summaries.

6. Conclusion and Further Work

We hope that the resource presented in this article will prove to be valuable for the evaluation of Polish summarizers and can also be used for studies regarding the nature of the human summarization process of news texts written in Polish language.

Both parts of the corpus (extractive and abstractive) will be also used to evaluate the difference of readability (e.g. with standard Gunning's FOG method (Gunning, 1968) or its Polish equivalent, Pisarek's method (Pisarek, 1969)) in manually and automatically created summaries. Our hypothesis is that summaries, with their urge to convey as much information as possible using limited number of words, are probably more complex and therefore harder to understand by the reader than the original text.

Lack of such resource was a major obstacle in comparing existing summarization resources and applicability of summarization evaluation methods to Polish language. Its free word order may prove to be a major obstacle in using the popular ROUGE (Lin, 2004) metric for the latter purpose, as equally informative summaries with a different word order would receive different scores, as the basic ROUGE metric is based on n-gram co-occurrences.

Acknowledgements

The work reported here was co-funded by the European Union from resources of the European Social Fund, project PO KL "Information technologies: Research and their interdisciplinary applications" as well as within the Applied Technology for Language-Aided CMS project (ATLAS), Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467).

7. References

Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kopeć, M., and Vertan, C. (2013). Summarizing short texts through a discourse-centered approach in a multilingual context. In Neustein, A. and Markowitz, J. A., editors, *Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems*. Springer-Verlag, Heidelberg/New York.

Branny, E. and Gajęcki, M. (2005). Text Summarizing in Polish. *Computer Science*, 7:31–48.

Ciura, M., Grund, D., Kulików, S., and Suszczańska, N. (2004). A System to Adapt Techniques of Text Summarizing to Polish. In Okatan, A., editor, *International Conference on Computational Intelligence*, pages 117–120, Istanbul, Turkey. International Computational Intelligence Society.

Dudczak, A., Stefanowski, J., and Weiss, D. (2008a). Automatic selection of sentences for Polish newspaper articles (Automatyczna selekcja zdań dla tekstów prasowych w języku polskim, in Polish). Technical Report RA-03/08, Institute of Computing Science, Poznań University of Technology, Poland.

Dudczak, A., Stefanowski, J., and Weiss, D. (2008b). Comparing Performance of Text Summarization Methods on Polish News Articles. In *Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference*, pages 249–258, Zakopane, Poland.

Dudczak, A., Stefanowski, J., and Weiss, D. (2010). Evaluation of Sentence-Selection Text Summarization Methods on Polish News Articles. *Foundations of Computing and Decision Sciences*, 1(35):27–41.

Dudczak, A. (2007). Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego (En. Application of selected data exploration methods to summarization of Polish newspaper articles). MSc thesis.

Gunning, R. (1968). *The Technique of Clear Writing*. McGraw-Hill.

Głowińska, K. and Głowiński, C. (2003). Summarization of Polish texts. In Goźdź-Roszkowski, S., editor, *The proceedings of Practical Applications in Language and Computers (PALC 2001)*, volume 7, pages 193–206. Peter Lang.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nenkova, A., Passonneau, R. J., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2).

Ogrodniczuk, M. and Karagiozov, D. (2011). ATLAS — The Multilingual Language Processing Platform. *Procesamiento del Lenguaje Natural*, 47:241–248.

Ogrodniczuk, M. and Lenart, M. (2013). A Multi-purpose Online Toolset for NLP Applications. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S., editors, *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*, pages 392–395. Springer Verlag, Berlin, Heidelberg.

Pisarek, W. (1969). Jak mierzyć zrozumiałość tekstu (En. How to measure text readability). *Zeszyty Prasoznawcze*, 4:35–48.

Presspublica. (2002). Korpus Rzeczpospolitej. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.

Rath, G. J., Resnick, A., and Savage, R. (1961). The formation of abstracts by the selection of sentences. *American Documentation*, 12(2).

Rotem, N. (2003). The Open Text Summarizer. [on-line] `http://libots.sourceforge.net/`.

Suszczańska, N. and Kulików, S. (2003). A polish document summarizer. In Hamza, M. H., editor, *Applied Informatics*, pages 369–374. IASTED/ACTA Press.

Świetlicka, J. (2010). Metody maszynowego uczenia w automatycznym streszczaniu tekstów (En. Machine learning methods in automatic text summarization). MSc thesis.