World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:1, No:8, 2007

A System to Adapt Techniques of Text Summarizing to Polish

Marcin Ciura, Damian Grund, Sławomir Kulików, and Nina Suszczańska

***Abstract*—**This paper describes a system, in which various methods of text summarizing can be adapted to Polish. A structure of the system is presented. A modular construction of the system and access to the system via the Internet are signaled.

***Keywords*—**Automatic summary generation, linguistic analysis, text generation.

I. Introduction

THIS article describes a fragment of our system for automatic text summarizing. As there already exist several well developed methods of summarizing (i.e. selecting essential facts), we began with adapting them to Polish. The described system is a skeleton, to which various summarizing methods can be plugged. We assume that these methods weigh each sentence basing on the results of a linguistic analysis. The system then generates a summary by selecting the most important sentences. We concentrated on the development of a text analyzer and a text generator. We thus supplied an interface, which can be used to test summarizing methods and to adapt them to Polish.

To construct the system we used the Linguistic Analysis Server [1] and mechanisms that allow for remote processing via the Internet. The system can be used as a standalone application, but its part is also used in a system of translation from a Polish written text to the Polish Sign Language [2].

II. The Structure of the System

Figure 1 contains a scheme of summary generation. The system consists of two main components. The first component is the Linguistic Analysis Server (LAS), used during text analysis. The second component is the PolSum2 application (Polish Summarizer, version 2), generating summaries from selected essential sentences. It processes the output of LAS. These components are described in next sections.

III. LAS

The Linguistic Analysis Server is a multi-purpose system, performing various kinds of text analysis. In PolSum2 we use only one of these kinds. The Linguistic Analysis Server takes a source text as input and outputs additional linguistic information about it in XML format. To communicate with the server we can use SOAP (Simple Object Access Protocol). It is a platform independent standard, so we can use the server from any operating system.

In addition, the server is accessible via WWW pages at http://thetos.zo.iinf.polsl.gliwice.pl/las/. The stages of the analysis are described below.

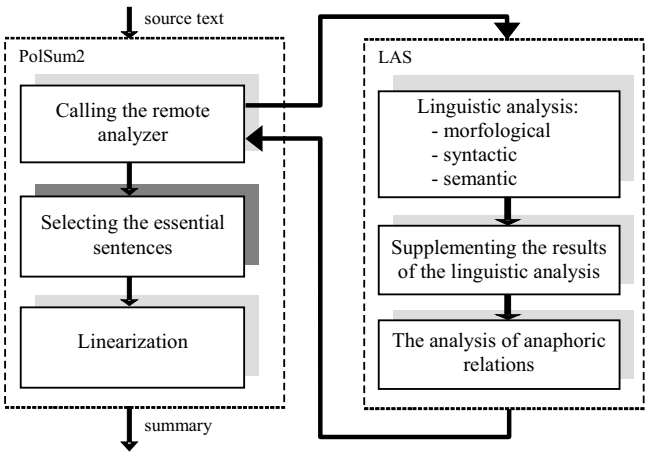

FIG. 1. THE STAGES OF SUMMARY GENERATION

A. Linguistic analysis

Linguistic analysis is divided into three stages. The morphological analysis segments the source text into words, and determines the type and the morphological features of each word. The syntactic analysis builds a parse tree on the basis of Syntactic Group Grammar for Polish (SGGP) and finds syntactic relations between groups. SGGP divides compound sentences into simple sentences. These stages were described in our previous papers; a detailed explanation of them can be found in [3]-[5]. The semantic analysis determines the semantic roles of each syntactic group. To this end, it uses the information about the semantic environment of a predicate (a verb). A dictionary of verb valence [6] contains syntactic schemata for verbs. Using them the system can determine, for a given meaning, the requirements for syntactic groups related to the verb. The analyzer uses the dictionary and adjustment rules in its search for syntactic groups occupying proper places in the verb schema. For example, after the syntactic analysis of the sentence 'Nieroztropność młodzi ludzie przypłacają zdrowiem' (lit.: 'Unwisdom young

M. Ciura, D. Grund, S. Kulikow, and N. Suszczanska are with the Institute of Informatics, Silesian University of Technology, 44-100 Gliwice, Poland (e-mails: mciura@star.iinf.polsl.gliwice.pl, dgrund@poczta.fm, skulikow@star.iinf.polsl.gliwice.pl, nsuszcz@star.iinf.polsl.gliwice.pl).

people they-pay with-health', 'The youth pay for their unwisdom with their health'), we get four syntactic groups:

- NG1 – nieroztropność (unwisdom)
- NG2 – młodzi ludzie (young people)
- NG3 – zdrowiem (with-health)
- VG1 – przypłacają (they-pay)

All these groups are combined to a sentence S1 = {NG1,NG2,NG3,VG1}, based on the verb group VG1. For VG1 we find in the dictionary a schema: Ngn1+VG1+Ngac-c2+NGi3. Then we assign the syntactic groups to the schema. There is a result:

- subject: NG1 or NG2
- predicate: VG1
- object in accusative case: NG1
- object in instrumental case: NG3

There are two candidates to be a subject, but NG1 is the only one candidate to be an object in accusative case, so NG2 is the subject. In this case a full semantic analysis of syntactic groups is not required.

B. Supplementing the results of the linguistic analysis

Missing elements of a syntactic schema of a sentence are determined basing on the results of the semantic analysis. Either a subject, an object, or a predicate can be missing. If necessary, the results of the analysis are supplemented by a predicate group (VG) or a subject or object group (NG). Syntactic schemata are also used here. After assigning the syntactic groups to the schema, the schema can be partially filled. This situation occurs when some syntactic group from the sentence is an anaphora. We can read syntactic features, semantic features, and a syntactic role of the anaphora from the schema. Then the analysis of anaphoric relations is more precise. If the anaphora is an ellipsis, then we add a proper pronoun as a subject or an object. If a predicate is missing, then we add the verb 'być' ('to be') in a proper form. The algorithm of supplementing the missing words is not so sophisticated. In a future version we plan to improve it, using an analysis of ellipses.

C. The analysis of anaphoric relations

The analysis of anaphoric relations is divided into three stages: searching for an anaphora, searching for a target of the anaphora, and determining how to substitute the target for the anaphora. As a result, we get information about sentence interconnections, and how to keep these connections, i.e. which form of the target has to be substituted for an anaphora.

The problem of anaphoric relations analysis [7]-[9] is complex, so we focused only on some subset of anaphora. At present we take only pronouns and some conjunctions [10] into account. It should be noticed that a missing subject is supplemented by a pronoun in a previous stage (supplementing the results of the linguistic analysis). We do not take into account such anaphoras that repeat some notion, which occurred earlier in the source text. To take into account these anaphoras we should use a database of term dependency, for example thesaurus.

We made assumptions while searching for anaphora targets (antecedents): we look only backward in the scope of two simple sentences (we do not look forward), we take into account only separate words. Since an anaphora can target to another anaphora, so indirectly the scope is wider than two sentences. We select an antecedent from candidates using the following criterion:

- it must belong to a noun group (NG)
- its gender must match with the gender of anaphora
- if it is an anaphora, then it must be joined with an antecedent

If there are still some candidates left, then we select from them exactly one. A function of a syntactical group containing the antecedent is the main criterion of selecting the antecedent. The preference is as follows (the most important is the first):

- subject function
- other function (for example object)
- no function

As a result it returns the information how to substitute the antecedent for the anaphora. This information can be used by the system or not. To perform this substitution, we take the base form and the type from the antecedent, but the features from the anaphora. If the anaphora features and the antecedent features differ, then we have two solutions: generate a proper form of the antecedent or leave the base form. Since Polish is an inflected language, we try to generate a proper form of antecedent. In this way the generated text is more readable by human.

To generate a proper form of antecedent we use morphological database [11]. The morphological database contains a morphological description of over 50 thousand Polish words, including over 30 thousand nouns (presently only nouns take part in anaphora resolution). It is stored in an acyclic finite-state automaton that represents the relation between text forms and words with their grammatical attributes. While the database recognizes variant word forms (which occur in 1 Polish noun out of 8) when used to analyze texts, it outputs only the most prevalent variant form when used to generate texts.

The use of this database required creating an interface layer because the set of grammatical attributes used in the database differs from that used in the system. Specifically, all requests are translated from a 3-gender, 7-case system into a 4-gender, 8-case one. Also, the system distinguishes the nouns according to their nominative form, while the database takes into account the whole declension paradigm. The format of the request is: (a noun in nominative, its gender [masculine, feminine or neuter], its number [singular or plural], its case [nominative, genitive, dative, accusative, instrumental, locative, vocative]).

Example nouns that decline differently depending on their meaning are: 'pilot' ('pilot' or 'remote control'), 'rząd' ('government' or 'row'). For such nouns a set of pairs (possible form, internal code of the noun) is generated. To distinguish

them we try to identify them on morphological level. We generate an antecedent form using its features and then we search for a matching candidate. Such method was used in example in Figure 2.

Since we use only the morphological level, we can distinguish the meaning of words only in the same case. Usually they are distinguishable in genitive case and accusative case. Since the method of searching for an anaphora or for an antecedent is a heuristics, it is possible to make a mistake. To improve it, we should use semantic features. Among such features we should mainly use semantic gender, which sometimes is different than morphological gender.

IV. PolSum2

PolSum2 is a new version of the PolSumm application [12]. We focused on providing mechanisms that facilitate adapting various summarizing methods. We thus developed a test bed for these methods. The stages of summarizing are described below.

A. Calling the remote analyzer

In this stage we only communicate with the LAS server. The source text is sent to the server, and the results of the analysis are received.

B. Selecting the essential sentences

In this stage the essential sentences are identified. These sentences will be selected to a summary. Researchers are provided with a coherent environment to implement and test their methods.

The essential sentences have a high weight, computed for each simple sentence from the results of source text analysis. Since the anaphoras are substituted, we can select among more sentences, because any text composed of them is coherent. Otherwise we would have to select among compound sentences or blocks of sentences to keep the resulting text coherent. In this way the building blocks of the summary are smaller than sentences (we do not need to select all compound sentences). But there are some disadvantages to this approach. We sometimes do not analyze all the interconnections between the sentences, so the generated summary can consist of sentences with broken connections.

We assume that the user chooses how many essential simple sentences should be selected to the summary, as a numeric range. These sentences are copied (sometimes with modifications) from the source text to the summary, preserving their order.

C. Linearization

In this stage a textual form of each sentence is generated. Text generation is divided into two threads. The first thread is a generation of proper form of words. The second one is a generation of sentence, in other words to place generated words in proper position in the sentence.

Generation of words is used only for antecedents in case of substitution (other words are already in proper form in the source text). In this stage we only use results from the analysis or anaphoric relations, since there is given a way of substitution.

In addition we reduce homonyms for the same word, where various variants appear during morphological analysis. When we generate text we may lose morphological information.

Information about substitution for anaphora need not be used when two simple sentences with an anaphoric relation were selected to the summary. We may leave this relation, so the generated summary is “more natural” for the reader. In some cases it is possible that these simple sentences become transformed into a compound sentence.

We assumed that we do not change the order of words, so to generate text of the summary we place generated words in the same order as they appear in the source text.

Figure 2 shows an example of text generation. The source text consists of three sentences with anaphoric relations (with use of pronouns). Moreover, an example of identifying the word on the morphological level (described in section III C) is presented there. The base form of the word 'rzędzie' (row) is 'rząd' ('row' or 'government'), but after identifying the word we use the form 'rzędu' (row) not 'rządu' (government) for substitution for anaphora. After text generation with substitution for anaphora we get sentences that can exist separately without significant loss of meaning.

Ja jestem w rzędzie. On jest daleko. Mój przyjaciel idzie do niego.
(I am in a row. It is far away. My friend goes to it.)

Ja jestem w rzędzie. Rząd jest daleko. Mój przyjaciel idzie do rzędu.
(I am in a row. The row is far away. My friend goes to the row.)

FIG. 2. EXAMPLE OF TEXT GENERATION WITH SUBSTITUTION FOR ANAPHORA

V. Conclusion

This paper presents a system which helps to adapt summarizing methods for Polish and to test them. It is a prototype, so some improvements are needed. We should take into account anaphoric relations via repetition of the same notion, and in general we should improve the heuristics of anaphora relations analysis and text generation. It should be noted that the system is built from modules, so any part of the system can be changed (and of course the part responsible for the selection of the essential sentences should be changed often). On the other hand the system has some disadvantages. The primary disadvantage is that it supports only these summarizing methods that are based on selecting sentences from the source text.

References

[1] S. Kulików. Implementation of Linguistic Analysis Server for Thetos – Polish Text into Sign Language Translator. Studia Informatica, vol. 24, no. 3 (55), Gliwice 2003, pp. 171-178 (in Polish)

[2] P. Szmal, S. Kulików. Support for Deaf People at Web Browsing. 3rd IASTED International Conference Artificial Intelligence and Applications AIA'2003, Benalmadena, Spain 2003, pp. 13-17

[3] N. Suszczańska, M. Lubiński. POLMORPH, Polish Language Morphological Analysis Tool. 19th IASTED International Conference Applied Informatics AI'2001, Innsbruck, Austria 2001, pp. 84-89
[4] N. Suszczańska, P. Szmal, J, Francik, Translating Polish Texts into Sign Language in the TGT System. 20th IASTED International Multi-Conference Applied Informatics AI'2002, Innsbruck, Austria 2002, pp. 282-287
[5] P. Szmal, N. Suszczańska. Selected Problems of Translation from the Polish Written Language to the Sign Language. Archiwum Informatyki Teoretycznej i Stosowanej 13(1), 2001, pp. 37-51
[6] D. Grund. Komputerowa implementacja słownika syntaktyczno-generatywnego czasowników polskich. Studia Informatica, Vol.21, No 3 (41), 2000, pp. 243-256 (in Polish)
[7] M. Dalrymple. The Syntax of Anaphoric Binding. Stanford, CSLI Publications, 1993. Available: http://csli-publications.stanford.edu/
[8] R. Muskens. Categorial Grammar and Lexical-Functional Grammar. Proceedings of the LFG01 Conference. Stanford, CSLI Publications, 2001
[9] Y. W. Romanyuk. Grammatical Processes of Text Compression. Ph. D. Thesis, Institute of Linguistics NAN Ukraine, Kyiv 1996 (in Ukrainian)
[10] S. Kulików, J. Romaniuk, N. Suszczańska. A syntactical analysis of anaphora in the Polsyn parser. Proceedings of the International IIS:IIPWM'04 Conference, Zakopane 2004, Poland, pp. 444-448
[11] M. Ciura. Rozpoznawanie wyrazów w tekście polskim z zastosowaniem acyklicznych automatów skończonych. [Recognizing words in Polish texts with acyclic finite-state automata], PhD thesis, Silesian University of Technology, Faculty of Automatic Control, Electronics, and Computer Science, 2004.
[12] N. Suszczańska, S. Kulików. A Polish Document Summarizer. 21st IASTED International Conference AI'2003, Innsbruck 2003, pp. 369-374