Automatic summary evaluation based on text grammars

Emilia Branny
Aksis, University of Bergen
Allegaten 27
N-5007, Bergen, Norway
004855286177
emilia.branny@interia.pl

ABSTRACT
In this paper, I describe a method for evaluating automatically generated text summaries. The method is inspired by research in text grammars by Teun Van Dijk. It addresses a text as a complex structure, the elements of which are interconnected both on the level of form and meaning, and the well-formedness of which should be described on both of these levels. The method addresses current problems of summary evaluation methods, especially the problem of quantifying informativity, as well as the problem of objective measurement of well-formedness of text. It is believed that the ideas from this research can contribute to evaluation methods for algorithms transforming complex meaningful entities into other complex meaningful entities (text, hypertext, sound, vision).

Keywords
summarizing, summary, evaluation, method, algorithm, intrinsic, T-grammar, proposition, informativity, corpus

1. INTRODUCTION
Automatic text summarizing belongs to advanced text processing tasks. Unlike many simple text processing tasks, it has several features which make it a specific kind of problem:

- It is a heuristic problem, which means that it has more than one acceptable solution
- The correctness of a solution is somewhat arbitrary, because the criteria are fuzzy
- An ideal solution can be approximated, which means that a summary which fulfils the criteria up to a certain degree can also be accepted as valid, although it will be rated lower than a summary which fulfils the criteria better (contains fewer errors, gives a fuller account of the original text etc.)
- Some sets of parameters do not have any good solution, e.g. it is impossible to reduce a typical cooking recipe to 30% of its original size and still keeping it useful.

It has to be added that automatic text summarizers usually produce summaries by rewriting selected sentences to the output. On one hand this makes the task of evaluation easier, because mistakes inside sentences are unlikely to appear. On the other hand many inter-sentential mistakes are usually caused by missing sentences. These mistakes affect the logic of the summary and can make it completely unreadable.

There are several methods for evaluation of automatic summarization. They have been named in (Hassel 2004). The main types of methods are:

- intrinsic methods
- extrinsic methods

Intrinsic methods concentrate on the summary itself, trying to measure its cohesion, coherence and informativity, usually in comparison with other summaries of the same text ("gold standard"). Extrinsic methods measure the usefulness of the summary in some task, for example relevance assessment or reading comprehension.

A popular example of an intrinsic method is the method used for evaluating ScandSum and its ported versions. A group of people are asked to choose the most important sentences from the text to form a summary. Then, the "golden" summary consisting of the most frequently chosen sentences is compared to the one generated by the computer (by counting the sentences that overlap). Also coherence is rated in the computer-generated summary (by asking the users to assess it).

This method has some advantages. The main advantage is that it is based on very clear criteria and once the summaries by a group of people are made, the result can be computed automatically.

However, the method has also serious disadvantages, among them:

- it does not take into account the fact that the quality of a summary depends on its informative content and not on its convergence with other summaries,
- although incoherence and mistakes in surface structure can be tracked and measured very precisely, the method does not take into account the findings in this field, leaving the assessment of summaries solely to the "natural competence" of language users,
- it does not track misinformation, which means that if a summary gives wrong information (for example by accidentally creating false pronominal

interdependencies), this summary can still be rated very high.

These problems are general problems of intrinsic summary evaluation methods, which usually lack the tools for precise measurement of the summaries' features and depend on convergence with other summaries to measure informativity.

Extrinsic methods can track misinformation and measure informativity more precisely, but only in relation to a certain task. They do not account for coherence, cohesion, grammatical mistakes etc. unless they significantly influence the meaning.

This is the reason why I would like to propose a different method for summary evaluation. It belongs to intrinsic methods and is based on three criteria:

- informative content
- misinformation
- T-grammaticality

I believe that presented approach to text summarization can contribute to the understanding and evaluation not only of text summaries but also of other heuristic text processing algorithms which deal with complex and heavily interconnected content. The most important contribution intended here is to show how one can address the evaluation of meaningful entities, produced by a machine. In the future this approach can be adapted and applied to machine-created hypertext and hypermedia.

2. THEORETICAL BACKGROUND

The method is inspired by the theory of text by Teun Van Dijk. Therefore, some terminology from this theory will be necessary to understand the ideas behind the method. Let me shortly introduce these concepts.

The first idea which will be taken from the theory is the idea of a **T-grammar**, as presented in (Van Dijk 1972). A T-grammar is a way of describing a valid text structure in a formal way. The general task of a T-grammar is the formulation of the rules forming semantic structures and relating them to phonological structures in all the valid texts of a language. A T-grammar should take into account two levels of text structure:

- surface structure (relations between subsequent sentences, both grammatical and semantic)
- deep structure (textual macrostructures and superstructures which account for the structure of the text as a whole)

This has been the inspiration behind the idea of T-grammaticality. A text will be called T-grammatical if it can be produced by a T-grammar, thus it is a valid and well-formed text of a language.

The second idea which has been taken from Van Dijk and only slightly modified to be useful in summary evaluation is the idea of a **proposition**. If semantic value of a sentence is to be discovered, it is necessary to formalize its meaning. This formalized meaning of a sentence is called a proposition. From propositions macropropositions can be derived, which usually represent more general meaning.

3. INFORMATIVE CONTENT

Evaluation of informative content in a summary consists of three phases:

1. Preparing an "information list" - list of propositions present in the original text
2. Marking these propositions which should be present in a summary
3. Evaluating the presence of the selected propositions in the summary

The information list (in fact a list of macropropositions expressed in natural language - not in logic symbols as Van Dijk does) is made according to several principles:

1. The information from a sentence should be split into "atomic" propositions:

   "The Board of Governors, which comprises of 39 persons, sat yesterday for 3 hours."

   becomes

   "The Board of Governors comprises of 39 persons."

   "The Board of Governors sat for 3 hours."

   "The event took place yesterday."

2. Information which is presupposed or which is somehow implicated by text structure (present in the text but not fully explicit) should be included:

   "As we learn unofficially, Andrzej Wajda was honored by American Film Academy with an Oscar for lifetime achievement. The Board of Governors, which comprises of 39 persons, sat yesterday for 3 hours."

   produces

   "The Board of Governors is the organ which awards the Oscar on the behalf of American Film Academy."

   The relation between the Board of Governors and the process of award is implicated by the order of sentences. The theme (topic) of the second sentence must be something already referred to because it is in the position of theme, while the rest is the rheme.

3. Different degrees of generalization should be provided (by including derived, more general, macropropositions):

   "Let us also remind that Steven Spielberg has supported the Polish creator. Over 70 American artists, among them Woody Allen and Olivier Stone, have signed his letter".

   produces:

   "The candidature of Andrzej Wajda had vast support."

   "The candidature of Andrzej Wajda was supported by Spielberg."

   "The candidature of Andrzej Wajda was supported by Spielberg in a letter."

   "The letter by Spielberg was signed by many actors." (dependent on 16)

   "These were American actors" (dependent on 16 and 17)

"There were many of them." (dependent on 16 and 17)

"There were over 70 of them." (dependent on 16 and 17)

"The letter by Spielberg was signed by Woody Allen and Olivier Stone." (dependent on 16)

Note, that in cases where a piece of information does not make sense without some other information, the formulation of propositions can be dependent. Several propositions listed here are centred around the topic of a letter of support by Spielberg, their sense depending on the presence of proposition (16).

4. Obvious background information (which is the information presumably known by an average reader and not being a topic in the text) is omitted in the information list:

The propositions such as:

"A film has a producer."

"Oskar is awarded by American Film Academy."

"Wajda is Polish."

are not included in the information list.

Several educated users of the language are asked to perform the task of choosing these propositions from the information list, which should be definitely present in the summary. They are also asked to add a proposition of their own if they want to include it but did not find it in the list. Then the user responses are normalized according to a simple algorithm:

- if a proposition is marked and its generalization is not, the generalization gets marked
- if a proposition is marked and a proposition on which it is dependent is not marked, the latter gets marked

These actions are repeated on the set of propositions until the proposition list is fully normalized:

- for each proposition marked: all of its generalizations are marked
- for each proposition marked: all propositions it depends on are marked

After the normalization has been performed, the common part of the lists created by different users is chosen. It forms the basis for evaluation of informative content in the algorithm.

The informative content of the propositions, as provided by users, has to be quantified in order to measure the performance of the summarizer. This is not an easy task, because it is not clear how to establish the proportions between the propositions, especially if some of them are generalizations of others. The number of propositions is arbitrary, so it is not possible to check the presence of a proposition in a text and count the proportion of propositions which do and which do not appear in the summary. We could as well split the first macroproposition into several propositions such as: *"An Oscar was awarded to a Polish filmmaker", "An Oscar was awarded to Wajda", "An award was given to Wajda", "An international award of great significance was given to Wajda"*, just because we are able to generalize on the basis of commonplace knowledge. So, the assessment of informativity has to be performed in a different way than simply counting the number of propositions occurring in the summary.

That is the point where once again the text theory of Teun Van Dijk proved useful. In each of the analysed texts the focus changes in different parts of the text. We should be able to assign weight to the information on the basis of text structure and text focus. In order to do this we divide the propositions into disjoint groups, according to their focus. Each of these groups will have the same weight and the weight of single propositions will amount to the score of a group.

For the analysed text the following groups can be found in user-chosen propositions:

- unofficial information about the Oscar being awarded to Wajda
- vast support, including that of Spielberg
- "Pan Tadeusz" will be shown to the Pope (the users did not choose more general propositions!)
- getting an Oscar is not easy, Wajda himself does not expect it

Now, having these groups, one can check how far the information marked by users is present in the summary. This is done by checking if relevant propositions are present in the text. The general principle is that the score is binary for each proposition (the score of 1 means that a proposition is present and 0 means that it is not present).

However, sometimes not full information is found in the summary, but only a part of it, or its generalization. Then, part of the score can be assigned. We have to assess how the presence of a general proposition affects the probability of the proposition we want to include, or "how much" of it is present.

Also when a proposition is not explicit in the text but it can be deduced, the summary should be evaluated as having this macroproposition. The score may be lowered (multiplied by probability) if the transition is not clear. This is the case for example with these two propositions:

"The record was around 7000 km."

"They traveled from Halifax to Vancouver."

To derive the former from the latter, the user has to know where the places mentioned are situated and what the distance is between them. If the former proposition is in the users' list but the summary gives only the latter information, the score for this proposition should be 1, multiplied by the probability of the fact that the reader knows these facts or is able to check them quickly.

To conclude, I will give the formula to calculate the score.

M – score for a macroproposition m entailing all the propositions from one group

P – presence of single proposition p (binary score modified by probability)

I – total score for informativity

Macropropositions and propositions which make them up are numbered as follows:

$$m_1: p_{11}\ p_{12}\ p_{13}\ ..\ p_{1j}$$

$$m_2: p_{21}\ p_{22}\ p_{23}\ ..\ p_{2k}$$

…

m_n: p_{n1} p_{n2} p_{n3} .. p_{ny}

The formula for calculating the informativity of a summary is:

$$M_1 = (P_{11} + P_{12} + \ldots + P_{1j}) \quad / j$$

$$M_2 = (P_{21} + P_{22} + \ldots + P_{2k}) \quad / k$$

…

$$M_n = (P_{n1} + P_{n2} + \ldots + P_{ny}) \quad / y$$

$$\mathbf{I = (M_1 + M_2 + \ldots + M_n) \quad / n}$$

4. MISINFORMATION

The summary should be studied in comparison to the original text and misleading statements resulting from automatic processing should be detected. If there is false statement or false implication in the summary, the score of the summary must be lowered.

There can be different kinds of misinformation. The misinformation about who got an Oscar would be destructive for the summary. This would affect the main macroproposition so this would ruin all the other information. Information e.g. that Olivier Stone wrote a letter to support Wajda would also be misinformative but not to such extent because the main macropropositions in the text would be kept.

Misinformation is sometimes difficult to specify. The general rule is that if people deduce from the text some facts that are not true, and they cannot deduce the same facts from the original text, then this can be called misinformation. A somewhat funny example from an automatic summary is:

> *„Elderly people – they claim – are not so ill as most people think. One in two 75-year-olds is suffering from rheumatoid arthrisis and one in three has arterial hypertension, chronic heart failure or hearing deficiencies."*

The implication is that this data is not so bad and that it confirms the fact that elderly people do not have many illnesses. In the original text another sentence is between the two, which starts with "But…" and radically changes the meaning of the data. It cannot change statistical data, but it shows that the researchers' point of view is different that we could have expected from the summary. In the original text the statistics are shown as bad.

As to general rules for assessing misinformation, I suggest that misinformation should be tracked down in the summary, written down as propositions (obviously being false in the original text), and then a group of people should be asked to read the original text and the summary and to assess in what extent that misinformation ruins the message. This extent will be the score for misinformation. The text should get 0 for misinformation if they do not find misinformation or say that they do not get the point altogether (which means that it is T-grammaticality problem and not misinformation).

Misinformation will be marked as F.

F – degree of misinformation (falsehood) as assessed by the users.

For a completely false summary: F = 1

For a summary containing no misinformation: F = 0

5. T-GRAMMATICALITY

In a summary both the surface structure itself (pronoun references, causal and temporal relations between sentences etc.) and the way it is developed from macrostructures and superstructures should be subject to evaluation in. The summary should be subject to evaluation especially in its cohesion and coherence. Because it is possible to track the mistakes down and name them, this should be done for each sentence.

As an example let us examine the following summary:

> *"The Regional Court in Bialystok sentenced a teacher from Choroszcza, accused of accidentally causing death of two girls, to two years of imprisonment in suspense for 5 years.*
>
> *Moreover, it prohibited her to work at a teacher position or to accept a post connected with taking care of children and youth for three years. (1)*
>
> *Two sisters, Anna and Malgorzata B. (2)*
>
> *They asked the teacher for permission (3).*
>
> *Before going to the camp the children signed a set of regulations, in which one of the points categorically forbade leaving for the lake unaccompanied, however in fact this kind of events used to happen at the permission of the accused.*
>
> *Other supervisors kept discipline in their groups, although these were children from older classes.*
>
> *Piotr Sadziński"*

The most relevant information is present in the text, but it is difficult to read. I marked three places in the text where T-grammaticality mistakes can be tracked down.

When we read the text, the obvious information which should be given in point (3) is, what the permission referred to. This is a mistake in cohesion.

Another obvious mistake is at (2), where a sentence got broken. This can be seen as a malformed sentence with no verb: "Dwie siostry Anna i Małgorzata B" or "O zgodę" starting with capital letter can be interpreted as an orthographical mistake.

If we repair (2) and (3) in a simplest way, we get

> *"Two sisters, Anna and Malgorzata B. asked the teacher for permission to go to the water bank."*

Still, something is missing, There is too much new information in this sentence. The setting is not given and the sentence is difficult to read. There should be some introduction of the situation: place or settings. For example:

> *"The accident occurred at a school camp at Siemianówka Lake in Podlaskie Voivodship. Two sisters, Anna and Malgorzata B. asked the teacher for permission to go to the water bank."*

Still there is no explicit connection made between going to the water and death. However, one may expect that this connection can be made by the reader. The readers who were asked to answer if the connection existed answered "yes" without any hesitation and claimed that it was obvious.

As different mistakes have different impact on the text, they should be assigned different weights. Therefore the weight of a syntactical or ortographic mistake (referred to from now on as "minor mistake") will be 0,5, the weight of a mistake in text structure or cohesion (referred to from now on as "medium

mistake") will be 1 and the weight of coherence mistake (for each sentence involved) or a cohesion mistake where the agent is missing (referred to from now on as "major mistake") will be 2. If several mistakes appear in one sentence, the maximum score for the sentence will be 2.

So if S is a sentence:

x – number of minor T-grammatical mistakes in S

y – number of medium T-grammatical mistakes in S

z – number of major T-grammatical mistakes in S

T_S is the score of S for T-grammaticality.

$$(0{,}5*x + y + 2*z) < 2 \quad \Leftrightarrow \quad T_S = 0{,}5*x + y + 2*z$$

$$(0{,}5*x + y + 2*z) \geq 2 \quad \Leftrightarrow \quad T_S = 2$$

The scoring for T-grammaticality is related to the number of sentences in the summary. If the summary comprises of t sentences, then the score for T-grammaticality can be expressed by the following formula:

$$T = (\sum T_S) / 2t$$

6. TOTAL SCORE

In this evaluation system total score E is computed from the following formula:

$$\mathbf{E = I\,(1 - F)\,(1 - T)}$$

where:

I – informativity score

F – misinformation score

T – T-grammaticality score

7. RESULTS

The power of the proposed approach lies in the fact that the criteria of evaluation are not only precise and fairly objective, but also represent the natural expectations of humans towards a summary. In other intrinsic methods, based on artificial criteria such as the number of overlapping sentences, there will always be cases where the criteria are satisfied but the summary is not acceptable or poor. There will also be problems with evaluating the summaries against each other, where fine granularity of evaluation is needed.

During the research, the proposed method has been used on several text summarizers along with the overlapping sentences approach. It has proved to give more precise distinctions. There was even a case of two summaries which had the same proportion of sentences overlapping with human-produced extracts but their score by the new method differed quite a lot: it was 0,62 compared to 0,27. That happened because the informativity of these texts differs a lot. The details can be seen in a table below.

Table 1 Evaluation by the new method. Text 1751 from the automatically processed rzecz.iso corpus.

	Scores		
	People (average)	**SweSum, on English translation**	**SweSum Generic, on Polish text**
Proposition groups:			
The sentence	1	1	1
Circumstances of the accident	0,89	1	0
Justification of sentence	0,75	0,25	0
Informativity (I)	**0,88**	**0,75**	**0,33**
Misinforma-tion (F)	**0**	**0**	**0**
Score for T-grammatical mistakes ($\sum T_s$)	0	2,5	2,5
$(\sum T_s)/2t$ * I	**0,00**	**0,13**	**0,06**
Total	**0,88**	**0,62**	**0,27**

Table 2 Overlapping sentences in summaries. Text 1751 from the automatically processed rzecz.iso corpus.

	Percentage of…		
	Overlapping sentences with all human summaries	Overlapping sentences with one of human summaries	Original sentences
Average people*	**29%**	**62%**	**10%**
SweSum English	**14%**	**57%**	**29%**
SweSum Generic	**14%**	**57%**	**29%**

***Average people** - the results for human-produced summaries (overlapping or not in relation to other human-produced summaries).

As it can be seen, a comparison of the number of overlapping sentences does not give any information about the difference between the informativity of English SweSum and Generic SweSum.

8. CONCLUSIONS

The new method of evaluation could be particularly useful for benchmarking different summarizing algorithms. First of all, due to the fact that the approach is based on Van Dijk's theory of propositions, it allows to describe the content of the text independently of the exact shape of the sentences. This makes the method useful for evaluating not only the summaries consisting of selected sentences from the input text, but also those summaries in which there are new sentences generated by the algorithm.

Secondly, the research has shown that the method provides the necessary differentiation between better and worse summaries.

The method is relatively easy to use, although creating the propositions is time-consuming (as the list of propositions usually becomes longer than the text itself) and needs to be done by linguists. In order to benchmark automatic text summarizers, it is necessary to create a corpus of texts and rewrite it into propositions. All methods could then be tested on the same corpus.

Other tasks, namely choosing the most important propositions in the original text and finding them in a summary, are done by a group of educated language users. To make the process more efficient it is possible to implement a web application allowing to carry it out online.

9. ACKNOWLEDGMENTS

Special thanks to dr inż. Marek Gajęcki from AGH University of Science and Technology, Kraków and prof. K. De Smedt from University of Bergen. This research has been carried out at Multilingua Marie Curie EST Fellowship at Bergen, Norway.

10. REFERENCES

Branny Emilia (2005) "Text summarization in Polish", Master Thesis, Department of Computer Science, AGH University of Science and Technology in Cracow, Poland

Dalianis, H., Hassel M., de Smedt, K., Liseth A., Lech, T.C., Wedekind J. (2004) "Porting and evaluation of automatic summarization". In *Nordisk Sprogteknologi 2003. Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004*, edited by H. Holmboe (Museum Tusculanums Forlag), pp. 107-121, http://www.dsv.su.se/%7Ehercules/scandsum/ScandSumArsbog2003.pdf

Dalianis, H., Hassel M., Wedekind J., Haltrup D., de Smedt K., Lech, T.C. (2003) "Automatic text summarization for the Scandinavian languages". In *Nordisk Sprogteknologi 2002: Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004*, edited by H. Holmboe (Museum Tusculanums Forlag), pp. 153-163,http://www.dsv.su.se/~hercules/scandsum/ScandSumArsbog2002.pdf

Dalianis, H. (2000) *SweSum - A Text Summarizer for Swedish* http://www.dsv.su.se/%7Ehercules/papers/Textsumsummary.html

de Smedt, K., Liseth, A., Hassel, M., Dalianis, H. (2005) "How short is good? An evaluation of automatic summarization". In *Nordisk Språkteknologisk Forskningsprogram 2000-2004*, edited by H. Holmboe (Museum Tusculanums Forlag), pp 267-287, http://www.nada.kth.se/~xmartin/reports/ScandSum-yearbook2004-fullpage.pdf

Mazdak, N. (2004) "FarsiSum - a Persian text summarizer", Master thesis, Department of Linguistics, Stockholm University, http://www.dsv.su.se/%7Ehercules/papers/FarsiSum.pdf

Pachantouris, G. (2005) "GreekSum - A Greek Text Summarizer" Master Thesis, Department of Computer and Systems Sciences, KTH - Stockholm University. http://www.dsv.su.se/%7Ehercules/papers/GeorgePachantouris-MasterThesis.pdf

Van Dijk, T. A. (2004) *From Text Grammar to Critical Discourse Analysis. A brief academic autobiography. Version 2.0.* http://www.discourse-in-society.org/From%20Text%20Grammar%20to%20Critical%20Discourse%20Analysis%20-%202.htm

Van Dijk, T. A. (1988) *News as discourse* (Hillsdale, NJ: L. Erlbaum Associates)

Van Dijk, T. A. (1980) *Macrostructures : an interdisciplinary study of global structures in discourse, interaction, and cognition* (Hillsdale, NJ: L. Erlbaum Associates)

Van Dijk, T. A. (1972) *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics* (The Hague: Mouton)