Using semantic roles to improve summaries

Diana Trandabăț
Faculty of Computer Science
University "Alexandru Ioan Cuza" Iasi
Iasi, Romania
dtrandabat@info.uaic.ro

Abstract

This paper describes preliminary analysis on the influence of the semantic roles in summary generation. The proposed method involves three steps: first, the named entities in the original text are identified using a named entity recognizer; secondly, the sentences are parsed and semantic roles are extracted; thirdly, selection of the sentences containing specific semantic roles for the most relevant entities in text. Although the method is language independent, in order to check its viability, we tested the proposed approach for Romanian summaries.

1 Introduction

Text summarization refers to the task of shortening a long text. There are two major directions in text summarisation: the *extractive* and the *abstractive* paradigm (Mani, 2001). The first approach in creating summaries (most common) is based on identifying important words in texts by using their frequencies, and determining those sentences that contain a bigger number of important words. These sentences are extracted from the original text, and taken to constitute the summary. In this paradigm, the summarization is performed through sentence extraction: the summary is a subset of the sentences in the original text.

An alternative approach is to build a summary consisting of sentences that don't necessarily have to show up in that specific form in the source text. This requires a certain amount of deeper understanding of the text. This method can also be applied in the case of very large texts, such as a whole novel, where neither the determination of most significant sentences based on occurrences of frequent words, nor building discourse structures could be of help. In these cases, other methods, mainly expanding a collection of predefined flexible summary patterns (based for instance on the genre of the novel, or on some data on the main characters of the novel, a time and place positioning, and a rather shallow sketch of the initiation of the action) could be applied.

Our approach to summary building uses the first method, sentence extraction. However, the novelty of our approach consists in basing the extraction of different sentences from the original text on semantic role analysis, an association which is not yet explored at its full potential. The method is language independent, provided that named entity and semantic roles extraction modules are available.

The next Section introduces the sentence extraction phase of the summary generation using semantic roles. Section 3 presents the named entity recognition system use to identify entities in the initial text, while Section 4 presents the semantic role labeling procedure. The last section presents preliminary results obtained on 20 summaries, and discusses further development of the system.

2 Generating Summaries based on Semantic Roles

The natural language processing community has recently experienced a growth of interest in semantic roles, since they describe WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW etc. for a given situation, and contribute to the construction of meaning. If for text analysis, semantic roles have gained their way into natural

language analysis systems (see for instance Lluis et al., 2008; Surdeanu et al., 2003), they are rarely used at their full potential for text generation.

Christopherson (1981) was among the first to investigate the usefulness of semantic roles in summaries. More recently, Suanmali et al. (2010) used semantic roles and WordNet (Fellbaum, 1998) to compute the semantic similarity of two sentences in order to decide if the sentences are to be kept or not in the summary. The proposed method is a further step in this direction, combining semantic roles and named entity for sentence extraction.

The overall pipeline architecture of the proposed method is presented in Figure 1.

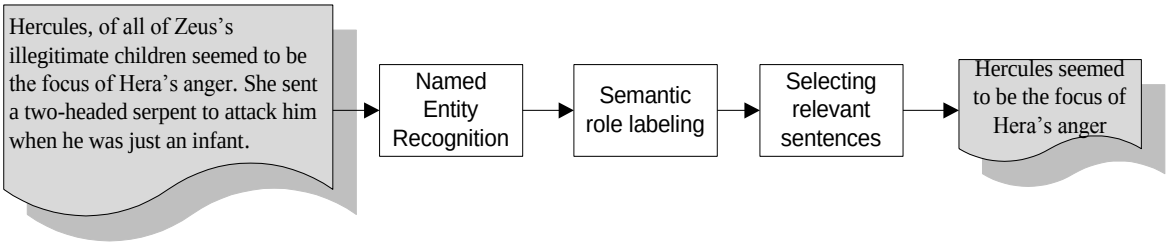

Figure 1. Overall architecture of the proposed summary generation method based on semantic roles

The method presented in this paper works in three steps: first, the original text is parsed for named entities; secondly, semantic roles are extracted from the sentences containing named entities; thirdly, sentences are selected to be kept in the summary, based on the semantic role the named entity has. Each module is detailed in the Sections below.

2.1 Identifying entities

In order to identify the semantic role a specific entity express, the entity must be first identified in the text. This is the task of named entity recognition (NER). NER systems typically use linguistic grammar-based techniques or statistical models (an overview is presented in (Nadeau and Satoshi Sekine. 2007)). Hand-crafted grammar-based systems typically obtain better precision, but at the cost of lower recall and months of work by experienced computational linguists. Besides, they are hard to adapt to new domains. Statistical NER systems typically require a large amount of manually annotated training data. Machine learning techniques, such as the ones discussed in (Scurtu et al., 2009) or (Nadeanu, 2007), allow systems to be adapted to new domains and perform very well for coarse-grained classification, but require large training data.

Thus, as a preprocessing module for our summary generation system, we used a Named Entity Recognition component for Romanian, based on linguistic grammar-based techniques and a set of resources. The NER system is based on two modules, the named entity identification module and the named entity classification module. After the named entity (NE) candidates are marked for each input text, each candidate is classified into one of the considered categories, such as Person, Organization, Place, Country, etc.

The major drawback of the sentence extraction approach for summaries generation is that it ignores the referential expressions that could occur in the initial text and should have been kept in the summary. Thus, due to the elimination of previous sentences, their antecedents may not be present anymore, resulting in incomprehensive readings. For example, consider the following text to be summarized:

Hercules, of all of Zeus's illegitimate children seemed to be the focus of Hera's anger. She sent a two-headed serpent to attack him when he was just an infant.

The summary of this very short fragment, using the sentence elimination method, could (hypothetically) be:

She sent a two-headed serpent to attack him.

which is really incomprehensible if no explanation is provided of who is "*she*" or "him".

One way to increase the coherence of such summaries is to derive first the discourse structure of the text and to guide the selection of the sentences to be included into the summary by a score that considers both the relevance of the sentence in a discourse tree and the coherence of

the text[1], as given by solving anaphoric references. For the summary example above, solving anaphoric references means identifying "*she*" as *Hera* and "*him*" as *Hercules*. Thus, the provided summary becomes readable:

Hera sent a two-headed serpent to attack Hercules.

Therefore, after identifying named entities and their types (person, organization, place, etc.), a simple anaphora resolution method, based on a set of reference rules, is applied to our input text, in order to link all entities to their referees.

The anaphoric system we used is a basic rule-based one, focusing on named entity anaphoric relations. Thus, we developed a rule-based system that performs the following actions:

- identifies a subset of a named entity with the full named entity, if it appears as such in the same text. For instance, *Caesar* is identified with *Julius Caesar* if both entities appear in the same text. Similarly, the *President of Romania* and the *President* are considered anaphoric relations of the same entity, if they appear in a narrow word window in the text.
- solves acronyms using a gazetteer we have initially built over the Internet, and which is continuously growing in size. For instance, *United States of America* and *USA* are co-references.
- searches for different addressing modalities and matches the ones that are similar. For instance, *John Smith* is co-referenced with *Mr. Smith*, and *Mary and John Smith* is co-referenced with *The Smiths*, or *The Smith Family*.
- solve pronominal anaphora in a simplistic way. Thus, if a pronoun (i.e. *she, he, him, his* etc.) is found in the text, and in the preceding sentence a named entity with the entity type person is found, then we create an anaphoric link between the pronoun and its antecedent. A similar rule exists for companies, where the pronoun *it* may be linked to *the Insurance Company*, for instance. Lists stating these correspondences are presently used and, although the rules are limited so far, our tests show that the overall accuracy of the summarization system benefits from this simple anaphoric resolution system for named entities.

The next step is the identification of the semantic roles that each named entity plays.

[1] A detailed analysis of the coherence of different texts is presented in (Cristea and Iftene, 2011).

2.2 Identifying semantic roles

Fillmore in (Fillmore, 1968) defined six semantic roles: *Agent*, *Instrument*, *Dative*, *Factive*, *Object* and *Location*, also called *deep cases*. His later work on lexical semantics led to the conviction that a small fixed set of deep case roles was not sufficient to characterize the combinatorial properties of lexical items, therefore he added *Experiencer*, *Comitative*, *Location*, *Path*, *Source*, *Goal* and *Temporal*, and then other cases. This ultimately led to the theory of Frame Semantics (Fillmore, 1982), which later evolved into the FrameNet project[2].

In the last decades, hand-tagged corpora that encode such information for the English language were developed (VerbNet[3](Levin and Rappaport, 2005), FrameNet (Baker et al., 1998) and PropBank [4] (Palmer et al., 2005)). For other languages, such as German, Spanish, and Japanese, semantic roles resources are being developed. For Romanian, Trandabăț and Husarciuc (2008) have started to automatically build such a resource.

For role semantics to become relevant for language technology, robust and accurate methods for automatic semantic role assignment are needed. With the SensEval-3 competition[5] and the CONLL Shared Tasks[6], Automatic Labeling of Semantic Roles, identifying frame elements within a sentence and tag them with appropriate semantic roles given a sentence (Lluis et al., 2008), has become increasingly present among researchers worldwide. In recent years, a number of studies, such as (Chen and Rambow, 2003) and (Gildea and Jurafsky, 2002), has investigated this task on the FrameNet corpus. Role assignment has generally been modeled as a classification task. While using different statistical frameworks, most studies have largely converged on a common set of features to base their decisions on, namely syntactic information (path from predicate to constituent, phrasal type of constituent) and lexical

[2] FrameNet web page: http://framenet.icsi.berkeley.edu/
[3] VerbNet web page: http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html
[4] PropBank web page: http://verbs.colorado.edu/~mpalmer/projects/ace.html
[5] SemEval web address: http://www.senseval.org/
[6] ConLL web address: http://ifarm.nl/signll/conll/

information (head word of the constituent, predicate).

Semantic roles are classified in terms of how central they are to a particular verb. **Arguments** (or core semantic roles) instantiate required roles, which are in a close relation to the verb whose sense they complete, and **adjuncts** (or non-core semantic roles), which are more general roles that can apply to any verb.

Adjuncts represent circumstantial objects and can be of the following types: directions, locatives, temporal, manner, extent, reciprocals, secondary predication, purpose, cause, discourse, adverbials, modals, negation. For instance, temporal and locative adjuncts can be found in both sentences below:

John broke the window [at the school]$_{LOC}$ [yesterday]$_{TMP}$.

John visited his kids [at the school]$_{LOC}$ [yesterday]$_{TMP}$.

An important drawback in this domain is that most researches focus on text analysis, and text generation applications using semantic roles are not so well developed. In this context, using the semantic role labeling system presented in (Trandabat, 2010), we annotated the sentences containing entities from the input text with the semantic roles these entities play, and passed to the third step.

The semantic role system we used for Romanian was obtained by training 12 machine translation algorithms (see Trandabat, 2010) from the Weka framework (Hall et al., 2009) with different feature sets. After running all the classifiers for different modules (the module that separately identifies the semantic roles and classify them, or the module that jointly identifies semantic roles and classify them), their performance is compared, and the module that obtains the highest performance is considered the best configuration. The models for this best configuration are saved, and the best path is written to a configuration file. This configuration can then be used at a later time to annotate new texts with the developed SRL system.

The 10 fold cross-validation results of all classifiers are also saved since they provide a confusion matrix that can be used to see which classes were correctly predicted by different classifiers. The output of the system presented in (Trandabat, 2010) is a Semantic Role Labeling System, a sequence of trained models which can be used to annotate new texts..

2.3 Selecting relevant sentences

The third module of the summary generation system implies selecting, among the list of sentences from which summaries can be generated, the ones in which the entity has core semantic roles. The proposed method involves four main steps:

- Identifying the main character
- Extract sentences containing the main character
- Keep sentences with core roles for the specific character.
- Simplify sentences

There are two possible ways of *identifying the main character*: the easiest one is when the central character of the text is a-priori given as argument (in case a character-oriented summary is requested). Otherwise, the main character is considered to be the named entity having the higher number of occurrences in the text (including references, see Section 2.1). For the example below, the main character is considered to be Alcmene, with 9 occurrences.

Hercules was the son of Zeus and Alcmene. Alcmene's husband Amphiteryon was out avenging her brother's death at the hands of pirates. Zeus, disguised as Amphiteryon, came to her and told her stories of how he killed the pirates to avenge her brother's death. That night Zeus went to bed with Alcmene and impregnated her. The next day the real Amphiteryon returned with his stories of avenging the pirates, and he could not understand why his wife was irritated with him and seemed disinterested in the stories. It was then that Amphiteryon consulted a blind seer and became aware of what Zeus did.

For the *extraction of the sentences containing the main character*, both the entity as if, and its references, are considered. For the example above, the last sentence is kept out, as not containing the character or a reference to it.

The distinction between the situations when the *main character has core and non-core semantic roles* (or adjuncts vs. arguments) represents the backbone of our system. Thus, when the entity considered for the summary has a semantic role

that is mandatory for a sentence meaning (it is a core semantic role, such as an *Agent*), the sentence containing it is kept. In contrast, if a sentence contains the entity in a non-core position (expressing temporal, spatial, modal, etc. circumstances), then its meaning is not essential for the summary, and the sentence containing the entity will be discarded from the summary. As an example, in the sentence below, Alcmene (refered as *his wife*) is only part of a non-core semantic role (Content for the verb *understand*), so this sentence will be discarded and not kept for the final summary:

`The next day the real Amphiteryon returned with his stories of avenging the pirates, and [he]Cognizer [could not understand]TARGET [why his wife was irritated with him and seemed disinterested in the stories]Content.`

The last step involved a *simplification of the sentences*. This simplification is based on a set of heuristics using semantic roles. Thus, in a sentence, not only one verb requiring semantic roles may appear. In order to simplify these complex sentences, we only keep the predicate[7] for which the entity is a semantic role. To give an example, consider the sentence below:

`[Alcmene's]`$_{\text{Partner1}}$ `[husband]`$_{\text{TARGET}}$ `[Amphiteryon]`$_{\text{Partner2}}$ `was out avenging her brother's death at the hands of pirates.`

In this case, two predicates are annotated with semantic roles: *husband* as a relationship predicate (according to FrameNet), and *avenging* as an activity predicate. Simplifying this sentence means keeping only the semantic roles for the first predicate (husband), for which the main character plays a semantic role, i.e. keeping only "Alcmene's husband, Amphiteryon was out".

3 Discussion and Further Work

In this paper, we presented a summary generation system based on semantic roles. The main components of the system are dedicated to identifying named entities, marking semantic roles, and selecting the sentences of the text to be kept in the summary.

We evaluated the method on 20 summaries extracted from the *Legend of the Olympus* novel. In a first batch, 5 volunteers received full version of the 20 texts, and were asked to generate short summaries (about 10% of the size if the full text). A second batch of 5 volunteers received the initial text marked with semantic roles, and were instructed to create short summaries (the same 10%) using the semantic role information. Although the evaluation was only intended to give a feedback on the method, and a proper evaluation is still to be developed, the volunteers reported that knowing the semantic roles of entities and guiding the summary on it makes the summary generation task easier.

Acknowledgments

The research presented in this paper was funded by the Sectoral Operational Programme for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

References

Baker G., Collin F., Fillmore, Charles J., and Lowe, John B. 1998. *The Berkeley FrameNet project*. In Proceedings of the COLING-ACL, Montreal, Canada. 1998

Chen J. and O. Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003.

Christopherson, Steven L . 1981. *Effects of knowledge of semantic roles on summarizing written prose*. Contemporary Educational Psychology, Vol 6(1), Jan 1981, 59-65

Cristea and Iftene. 2011. If you want your talk be fluent, think lazy! Grounding coherence properties of discourse. Invited talk at SPED-2011, University of Brasov, May

Fillmore Charles J. 1968. The case for case. In Bach and Harms, editors, Universals in Linguistic Theory, pages 1-88. Holt, Rinehart, and Winston, New York, 1968.

Fillmore Charles J. 1982. *Frame semantics*, în Linguistics in the Morning Calm, Hanshin Publishing, Seoul , 1982, 111-137.

[7] In general, predicates are associated with verbs. However, semantic roles theories have recently accepted the existence of predicate-like nouns and adjectives, which can gather around them semantic roles, just like verbs do.

Gildea Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245-288, 2002

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1

Levin B. and M. Rappaport Hovav. 2005. Argument Realization. Research Surveys in Linguistics Series. Cambridge University Press, Cambridge, UK, 2005.

Mani Inderjeet, automatic summarization, John Benjamins Pub Co; ISBN: 1588110591 (hardcover), 1588110605 (paperback), 2001.

Marquez Lluis, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the Special Issue. *Computational Linguistics*, 34(2):145-159, 2008.

David Nadeau and Satoshi Sekine. 2007. *A survey of named entity recognition and classification*, Linguisticae Investigationes 30, no. 1, 3{26, Publisher: John Benjamin's Publishing Company

David Nadeau. 2007. *Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision*, PhD Thesis.

Palmer Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71-106, 2005.

Scurtu V., Stepanov E., Mehdad, Y. 2009. *Italian named entity recognizer participation in NER task@evalita 09*, 2009.

Suanmali, L., Salim, N. and Binwahlan, M. S.. 2010. SRL-GSM : *A Hybrid Approach based on Semantic Role Labeling and General Statistic Method for Text Summarization*. Journal of Applied Sciences 10(3) : 166-173

Trandabăţ D. 2010. *Natural Language Processing Using Semantic Frames*, PhD Thesis, University Al. I. Cuza Iasi, Romania

Trandabat Diana and Maria Husarciuc. 2008. *Romanian semantic role resource*. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.