Multi-Document Summaries based on Semantic Redundancy

Sanda M. Harabagiu*, **V. Finley Lăcătuşu, Steven J. Maiorano**
Department of Computer Science
University of Texas at Dallas
Richardson, TX 75083-0688
*sanda@cs.utdallas.edu

Abstract

This paper presents a technique for producing short summaries from multiple documents. This technique promotes the belief that informative short summaries may be generated when using conceptual representations of redundant semantic information, called *topic semantic signatures*. The results of DUC-2002 evaluations account for the advantages of using the techniques presented in this paper.

Introduction

One way of tackling the current textual information overload is by relying on summaries of documents that cover the same topic from multiple perspectives. Summaries compress the information content available in a long text or a text collection by producing a much shorter text that can be read and interpreted rapidly. At the core of automatic summarization techniques that produce coherent summaries stays the methodology of identifying in the original documents the relevant information that should be included in the summary. Similarly, Information Extraction (IE) is a technology that targets the identification of topic-related information in free text and translates it into database entries. Typically, IE systems extract around 10% of a document's textual content (cf. (Hobbs and et al.1997)). This represents a compression ratio that qualifies extraction techniques for multi-document summarization. Our automatic summarization technique builds on this observation.

To further progress in summarization and enable researchers to participate in large-scale experiments, the National Institute of Standards and Technology (NIST) has initiated in 2001 an evaluation in the area of text summarization called the Document Understanding Conference (DUC)[1]. For DUC-2002 NIST produced 59 document sets as test data. For this purpose NIST used the TREC disks employed in the question-answering track in TREC-9. Specifically these include articles from *Wall Street Journal (1987-1992)*, *AP newswire (1989-1990)*, *San Jose Mercury News (1991)*, *Financial Times (1991-1994)*, *LA Times* and *FBIS records*. Each set had between 5 and 15 documents, with an average of 10 documents. The documents were at least 10 sentences long, but there was no maximum length.

For the DUC-2002 evaluations, given a set of documents, four generic summaries had to be generated automatically, with lengths of approximately 200, 100, 50, and 10 words (whitespace-delimited tokens) or less. To generate such short summaries we have devised a method for creating semantic representations of the typical information for each topic. We have assumed that the most important information is identified in the most redundant information, called the *topic semantic signature*. These topic signatures are used to identify textual information that is extracted from documents to form the short summary. Additionally, the identification of the topic signatures in documents enables the ordering of extracted information in the summary.

The rest of the paper is organized as follows. Section 2 presents the ad-hoc extraction technique for producing topic semantic signatures from redundant information. Section 3 presents the multi-document summarization technique based on this redundant information whereas Section 4 reports and discusses the experimental results we obtained in DUC-2002. Section 5 summarizes the conclusions.

Ad-hoc Extraction of Redundant Semantic Information

The idea of representing the topic as a frame-like object was first advocated in the late 70's by DeJong (DeJong 1982), who developed a system called FRUMP (Fast Reading Understanding and Memory Program) to skim newspaper stories and extract the main details. The topic representation used in FRUMP is the *sketchy script*, which model a set of pre-defined particular situations, e.g. demonstrations, earthquakes or labor strikes.

[1]DUC is part of a Defense Advanced Research Projects Agency (DARPA) program, Translingual Information Detection, Extraction, and Summarization (TIDES), which specifically calls for major advances in summarization technology, both in English and from other languages to English (cross-language summarization)

Since the world contains millions of topics, it is important to be able to generate the sketchy script automatically from corpora. In addition some of the current large-scale lexico-semantic knowledge bases may be used to contribute with information for the generation of the topic signatures. In our methodology, we have employed WordNet (Miller 1995), the lexical database that encodes a majority of the English nouns, verbs, adjectives and adverbs.

Extracting Topical Relations from WordNet

WordNet is both a thesaurus and a dictionary. It is a thesaurus because each word is encoded along with its synonyms in a synonyms set called *synset*, representing a *lexical concept*. WordNet is a dictionary because each synset is defined by a gloss. Moreover, WordNet is a knowledge base because it is organized in 24 noun hierarchies and 512 verb hierarchies. Additionally WordNet encodes three meronym relations (e.g. HAS-PART, HAS-STUFF and HAS-MEMBER) between nouns and two causality relations (e.g. ENTAILMENT and CAUSE-TO) between verbs. Additionally we noticed that chains of lexico-semantic relations can be mined from WordNet to account for the connection between any pair of signature concepts of known topics. To illustrate how such chains of relations can be mined, we first consider two of the relations already encoded in WordNet and then show how additional relations can be uncovered as lexico-semantic chains between two concepts pertaining to the same topic. We call these lexico-semantic chains *topical relations*.

The sources of topical relations
In WordNet, a synset is defined in three ways. First it is defined by the common meaning of the words forming the synset. This definition relies on psycholinguistic principles, based on the human ability to disambiguate a word if several synonyms are presented. Second, the synset is defined by the attributes it inherits from its super-concepts. Third, a glossed definition is provided to each synonym. A GLOSS relation connects a synonym to its definition. We believe that glosses are good sources for topical relations, since they bring forward concepts related to the defined synset. We consider four different ways of using the glosses as sources for topical relations:

1. We extend the GLOSS relation to connect the defined synset not only to a textual definition but to each content word from the gloss, and thus to the synset it represents. For example, the gloss of synset {*bovine spongiform encephalitis, BSE, mad cow disease*} is *(fatal disease of cattle that affects the central nervous system; causes staggering and agitation)*. A GLOSS relation exists between the defined synset and *fatal, disease, cattle, affect, central nervous system, staggering* and *agitation*.
2. Each concept from a gloss has its own definition, and thus by combining the GLOSS relations, we connect the defined synset to the defining concepts of each concept from its own gloss.
3. The hypernym of a synset has also a gloss, thus a synset can be connected to the concepts from the gloss of its hypernym. Similarly to the IS-A relations, other WordNet lexico-semantic relations can be followed to reach a new synset and have access to the concepts used in its gloss. Such relations may include HAS-MEMBER, HAS-PART or ENTAILS and CAUSE-TO. Lexical relations based on morphological derivations, if available may be used too[2]. Morphological relations include the NOMINALIZATION relations, known to be useful in IE.
4. A synset can be used itself to define other concepts, therefore connections exist between each concept and all concepts it helps define.

Figure 1 illustrates the four possible sources of topical relations based on two of the WordNet relations, namely GLOSS and IS-A.

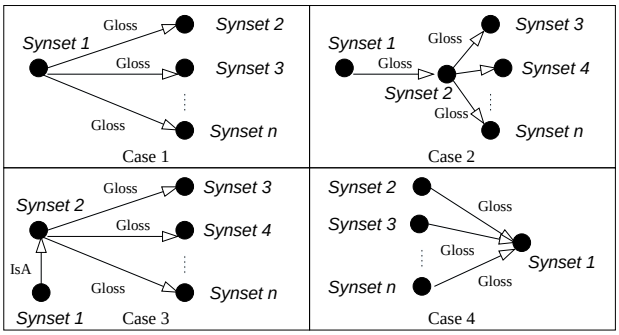

Figure 1: Four sources of topical relations.

Topical relations as Paths between WordNet Synsets
Two principles guide the uncovering of topical relations. First we believe that redundant connections rule out connections discovered by accident. Therefore, if at least two different paths of WordNet relations can be established between any two synsets, they are likely to be part of the representation of the same topic. Second, the shorter the paths, the stronger their validity. Consequently, we rule out paths of length larger than 4. This entails the fact that each topic may be represented by at least five synsets.

Figure 2 shows the topical relations produced by the paths originating at the WordNet synset {*mad cow disease*} and traversing concepts like {*mental illness*}, {*agitation*} or {*brain, mind*}. It is to be noted that each concept may be reached by at least two different paths of relations.

Topic Semantic Signatures

The representation of a topic can be viewed as a list of semantic roles, each role being a slot that is filled

[2]WordNet 2 already encodes derivational morphology.

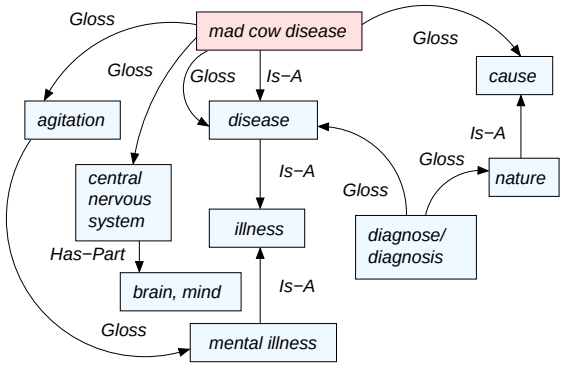

Figure 2: Topical relations for the "mad cow disease" topic.

by information extracted from text. The topical relations mined from WordNet have the advantage that they bring forward semantically-connected concepts deemed relevant to the topic. However these concepts cannot be mapped directly into a list of slots. First, WordNet was not devised with the IE application in mind - it is a general resource of English lexico-semantic knowledge. Because of this, some concepts relevant to a given topic may not be encoded in WordNet. Second, several WordNet concepts traversed by topical relations may be categorized under the same semantic role. Third, some semantic roles may be encoded in WordNet at a very abstract level, and thus they may never be reached by topical relations. Fourth, some of the semantic roles derived from topical relations may never be filled, since there is no corresponding information in the texts. To address all these issues, we have developed a corpus-based technique for creating ad-hoc lists of semantic roles for the topic signature representation for the collection. Our algorithm for ad-hoc topic semantic signature generation was inspired by the empirical approach for conceptual case frame acquisition presented in (Riloff and Schmelzenbach 1998).

Algorithm Ad-hoc Generation of Topic Semantic Signatures

Step 1: Extract all sentences in which one of the concepts traversed by topical relations is present. The concepts from the topical relations are used as seed lexical items for the identification of the signature slots.

Step 2: Identify all Subject-Verb-Object (SVO) +Prepositional attachments syntatic structures in which one of the topical concepts is used. For this purpose, we used the the Collins probabilistic parser (Collins 1996).

Step 3: Apply the COCKTAIL coreference resolution system (Harabagiu and et al.2001) and consider all the syntactic SVO structures involving all coreferring expressions of any of the nouns used in the syntactic structures discovered at Step 2.

Step 4: Combine the extraction dictionaries with WordNet to classify each noun from the structures identified at Step 2 and Step 3.

Step 5: Generate the semantic profile of the topic. For this reason we compute three values for each semantic class derived at Step 4: (1) **SFreq**: the number of syntactic structures identified in the collection; (2) **CFreq**: the number of times elements from the same semantic class were identified; and (3) **PRel** the probability that the semantic class identifies a relevant slot of the signature. Similarly to the method reported in (Riloff and Schmelzenbach 1998), **PRel** = **CFreq** /**SFreq**. To select the signature slots the following formula is used:
(**CFreq** $>$ *F1) or ((***SFreq** $>$ *F2) and (* **PRel** $>$ *P))*
The first test selects roles because of the semantic categories that are identified with high frequency, under the assumption that this reflects a real association with the topic elaboration in the collection. The second test promotes slots that come from a high percentage of the syntactic structures recognized as containing information relevant to the topic even though their frequency might be low. The values of *F1*, *F2* and *P* vary from one topic to another - we derive them from the requirement that a topic signature should not contain more than 5 slots.

Multi-Document Summarization

We decided to use topic semantic signatures in combination with coreference information resulting from the resolution of anaphors, e.g. pronouns or other referential expressions. Every time the topic signature would match against a text snippet, we would identify textual information to be extracted, called *topic snippet*. Thus for each topic signature T_i having the slots TS_i^1, TS_i^2, ..., TS_i^n we keep two additional forms of information: (1) the topic snippet $TextS_i^j$ that matched one of of its slots TS_i^j; and (2) all the entities from the text that corefer with the information filling any slot $TextS_i^j$. Figure 3 illustrates a snapshot of populated topic signatures and their mappings into topic snippets. The Figure illustrates some coreference chains as well. Coreference

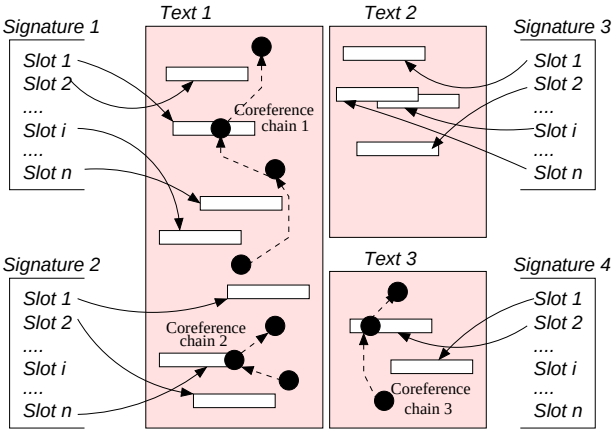

Figure 3: Mappings between topic signatures and topic snippets. Whenever a topic snippet contains an anaphor, pointers to all other entities with which it corefers are kept in a coreference chain.

information was obtained by using the COCKTAIL system (Harabagiu and et al.2001) on the texts.

To generate multi-document summaries we use two observations: (1) the order in which relevant text snippets appear in the original articles accounts for the coherence of the documents; and (2) to be comprehensible, summaries need to include sentences or sentence fragments that contain the antecedents of each anaphoric expression from relevant text snippets. Since all articles contain information about a given topic, it is very likely that a large percentage of the topic signatures share the same filler for one of the slots. In the case of the "natural disasters" topic, this filler was " *hurricane Andrew"*. We call this filler the *dominant event* of the collection. Additionally, we are interested in the topic snippets extracting information about other events that may be compared with the dominant event in the collection. Thus topic snippets are classified into three different sets: (a) S_1 - snippets about the dominant event; (b) S_2 - other snippets corresponding to a topic signature that has a slot filled by the dominant event; (c) S_3 - other topic snippets.

Multi-document summaries of length L are produced by extracting sentences from the document set in four different increments. The rationale for choosing four increments is based on the four different summary lengths imposed by the DUC evaluations, e.g. 50-word, 100-word, 200-word and 400-word long summaries. Since it is not known apriori how many topic snippets are extracted nor what is the cardinality of each S_i set, for each summary increment we perform at least one comparison with the target length L to determine if the resulting summary needs to be reduced or not. The multi-document summary is produced by the following algorithm:

Topic Signature-Based MD-Summarization (L)
Step 1: Select the most representative snippets.
To this end, for each topic signature TS_i from S_j, with $1 \leq j \leq 4$, for each slot TSS_i^j we count the frequency with which the same filler was used to fill the same slot of any other signature. The *importance* of TS_i is measured as the sum of all frequency counts of all its slots. This measure generates an order on each of the four sets of templates. Whenever there are ties, we give preference to the template that has the largest number of mapped text snippets traversed by coreference chains. Topic signature TS_0 is the most important element from S_1. If S_1 is null, the same operation is performed on S_2.
Step 2: Summary-increment 1.
Select sentences containing the text snippets mapped from TS_0 in the order in which they appear in the text from where TS_0 is selected. If anaphoric expressions occur in any of these sentences, include sentences containing their antecedents in the same order as in the original article.
if length(summary) $> L$ generate appositions for dates and locations and drop the corresponding sentences.
if length(summary) $> L$ drop coordinated phrases that do not contain any of the mapped text snippets.
while length(summary) $> L$ drop the last sentence.
Step 3: Summary-increment 2.
For each slot mapped into some other topic snippet from S_1, add its corresponding sentence/clause immediately after the sentence mapped by slot TS_0. If anaphoric expressions occur in any of these sentences, include sentences containing their antecedents in the same order as in the original article. Continue this process until either (1) the length of the summary is larger than $L-1$ or (2) there are no more sentences to be added.
Step 4: Summary-increment 3.
Repeat step 3 for snippet from S_2.
Step 5: Summary-increment 4.
Repeat step 3 for snippet from S_3.

Evaluation

In DUC-2002 multi-document summarization involved 59 document sets. For each test data set the multi-document summary generated by our system was compared with a gold-standard summary created by humans. For each data set, the author of the gold-standard summary assessed the degree of matching between the model summary and the summaries generated by the systems evaluated in DUC-2002. Each of these measures were scored on a scale between 0 and 4.

Q1: About how many gross capitalization errors are there?
Q2: About how many sentences have incorrect word order?
Q3: About how many times does the subject fail to agree in number with the verb?
Q4: About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) −causing the sentence to be ungrammatical, unclear or misleading?
Q5: About how many times are unreleted fragments joined into one sentence?
Q6: About how many times are articles (a, an, the) missing or used incorrectly?
Q7: About how many pronouns are there whose antecedents are incorrect, unclear, missing or come only later?
Q8: About how many nouns is it impossible to deterine clearly who or what they refer to?
Q9: About how many times should a noun or noun phrase have been replaced with a pronoun?
Q10: About how many dangling conjunctions are there ("and", "however" ...)?
Q11: About how many instances of repeated information are there?
Q12: About how many sentences strike you as in the wrong place because they indicate a strange time sequence, suggest a wrong cause−effect relationship, or just don't fit in topically with neighboring sentences?

Figure 4: Qualitative questions used to evaluate summaries in DUC-2002.

To compute the quantitative measures of overlap between the system-generated summaries and the gold-standard summary, the human-created summary was segmented by hand by assessors into *model units* (MUs), which are informational units that should express one self-contained fact in the ideal case. MUs are sometimes sentence clauses, sometimes entire clauses. In contrast, the summaries generated by the summarization systems were automatically segmented into *peer*

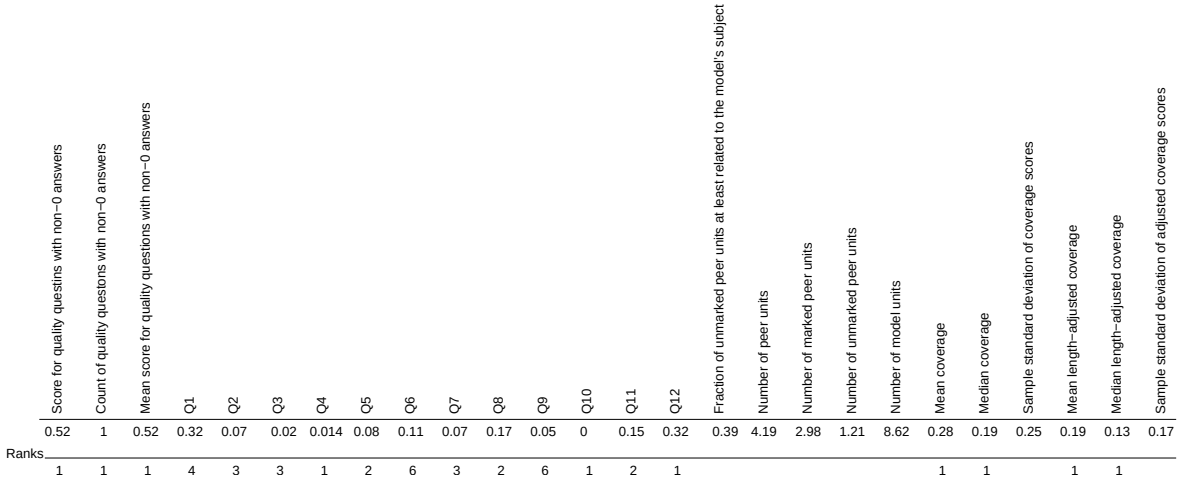

Figure 5: Results of the multi-document summarization evaluations in DUC-2002.

units (PUs) - which are always sentences. Figure 5 lists the results obtained for the multi-document summarization evaluations. The Figure also lists the results of the evaluations with respect to the accuracy with which the summaries responded the twelve questions listed in Figure 4. By ranking according to the mean coverage of PUs into MUs and the respective median coverage, the results of our methods were ranked on the first place. It also obtained the best rank for mean-length adjusted coverage and for median length-adjusted coverage.

For multi-document summaries, we considered also the *Precision* and *Recall* measures. Precision is calculated as the number of PUs matching some MU divided by the number of PUs in the peer summary, considering all summaries automatically generated for the same collection. We have obtained a precision of 20.66% and a recall of 20.70%. The precision was ranked as the third one whereas the recall was ranked as the first one among all the systems that participated in DUC-2002. As reported in (McKeown et al.2001), this estimate of the precision is conservative, since the number of PUs that are considered correct can be increased by considering information about the PUs not assigned to MUs.

Conclusions

In this paper we have shown that multi-document summarization of good quality can be obtained if topic signatures can be generated automatically. Our method of generating topic semantic signatures combined WordNet semantic information with redundancy information from the documents. We have presented a multi-document summarization procedure that incrementally adds information to create summaries of variable size. The decision of using incremental additions of sentences from multiple documents based on mappings into topic snippets produced very good results for coherence and organization in the DUC-2002 evaluations.

References

Michael Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL-96*, pages 184–191, 1996.

G. DeJong. An overview of the FRUMP system. In *Stategies for natural language processing*, W. Lehnert and M. Ringle Eds., pages 149-176, Lawrence Erlbaum Associates, 1982.

S. Harabagiu, R. Bunescu, and S. Maiorano. Text and Knowledge Mining for Coreference Resolution. In Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001), pages 55-62, 2001.

D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In *Finite State Language Processing*, edited by Emmanuel Roche and Yves Schabes, MIT Press, 1997.

K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman a nd S. Teufel. Columbia Multi-Document Summarization: Approach and Evaluation. In *Workshop Notes for the DUC-2001 Summarization*, pages 43–64, September 2001.

George A. Miller. WordNet: a lexical database for English. In *Communications of the ACM*, Vol.38, No.11:39–41, 1995.

E. Riloff and M. Schmelzenbach. An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixteenth Workshop on Very Large Corpora* , 1998.