A Generative Approach for Multi-Document Summarization using Semantic-Discursive information

Maria Lucía Castro Jorge, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos – SP
{mluciacj, taspardo}@icmc.usp.br

Abstract. *Multi-document summarization is the automatic production of a unique summary from a collection of texts. In this paper, we propose a statistical generative approach for multi-document summarization that combines simple information such as sentence position in the text and semantic-discursive information from CST (Cross-Document Structure Theory). In particular, we formulate the multi-document summarization task using a Noisy-Channel model.*

1. Introduction

Multi-Document Summarization (MDS) is the process of building a summary from a group of texts that have similar content (Mani, 2001).

In this work we explore a Generative Approach for MDS by using a Noisy-Channel framework (Shannon, 1948) for learning a MDS model. In this approach we integrate semantic-discursive knowledge to model different Multi-Document phenomena such as redundant, complementary and contradictory information. This semantic-discursive information across documents is given, for example, by CST model (Cross-Document Structure Theory) (Radev, 2000) and also RST (Rhetorical Structure Theory) (Mann and Thompson, 1987). This novel approach yields a theoretical generative learning model for MDS.

2. A Noisy-Channel approach for Multi-document Sumarization

The Noisy-Channel model is represented by a framework composed of three parts: a source, a noisy-channel and a decoder. This structure is showed in Figure 1.

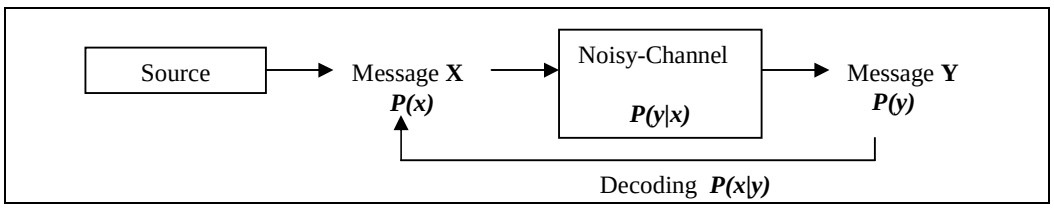

Figure 1. **Noisy-Channel Model**

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 224–228,
Cuiabá, MT, Brazil, October 24–26, 2011.

The source produces an original message which passes through a channel where some noise is introduced, and therefore, a corrupted message y is produced. The decoding stage consists in recovering the most likely **x** (original message), from a set of x's, given **y.** This whole process is formulated through the Bayes rule.

When instantiating MDS in the Noisy-Channel framework, we assume that the source will produce a multi-document summary. The probability for this summary is expressed by P(S) and it represents the language model of the summary which models factors such as grammaticality, coherence and cohesion in the summary. The probability of expanding the summary into a cluster (group of bigger texts from which the summary came from) is given by P(C|S). As an initial approach we will assume that this cluster will be a set of sentences, without making any difference to which texts of the cluster those sentences belong to. This two probabilities P(S) and P(C|S) are combined through the Bayes Rule to obtain P(S|C). In the decoding stage a set of possible summaries will instantiate the Bayes Rule for obtaining the best summary (1).

$$P(S \mid C) = \frac{P(C \mid S) \times P(S)}{P(C)} \quad (1)$$

In this work we will concentrate on the exploration of the channel model, P(C|S). For the moment, P(S) will be considered uniform among different clusters of texts. Similarly, P(C) will not be taken into account since it will be an observed value and, therefore, constant.

In the context of MDS, we consider that "noise" could be elements that emerge from multi-document phenomena factors such as redundancy, complementarity and contradiction. For instance, a sentence from a summary could generate complementary, redundant and contradictory sentences in the original texts. We use CST to model these factors by means of semantic relations among sentences of multiple documents. For example, complementary information can be modeled by the "Elaboration" relation; redundant information can be modeled by relations like "Equivalence", "Subsumption", *etc*.; and contradiction can be modeled by the "Contradiction" relation. Besides Cross-document relations, it can also be considered rhetorical information from RST. For this work, we will concentrate on a model based on Cross-document relations. We can formalize this generative process by establishing three initial conditions:

- A summary is a set of sentences SS={ SS_1, …,SS_n}
- The original texts from which each summary comes from form a cluster (group). This cluster is a set containing all the sentences of the original texts: CS= {CS_1, …, CSm}.
- A set of MDS phenomena factors is given, F={F_1, …,F_z}

To build this generative model, we consider having a parallel multi-document corpus containing clusters of texts annotated with CST relations, and their correspondent extractive summaries. Once we have the corpus available, P(C|S) is calculated by multiplying probabilities describing the chance of a sentence SSi to generate a quantity Nx of sentences through factor Fj. This is formulated in (2)

$$P(C \mid S) = \prod_{j=1} \prod_{i=1} P(N_x \mid SS_i, F_j) \quad (2)$$

The value of $P(Nx|SS_i,F_j)$ is obtained by dividing the number of summary sentences generating N sentences through factor F_j. by the total number of summary sentences SS_i in the corpus. A probability $P(F_j|SS_i)$ is associated to each probability, in order to express the chance of a summary sentence to be associated to the factor F_j. This is obtained dividing the number of sentences associated to F_j by the total number of cluster sentences CS_i generated by SS_i.

Another generative factor considered in our model is the location of the cluster sentences. For this, we associate a probability $P(N_y|SS_i,$ Location), which expresses the chance of SS_i generating a number N_y of sentences at a particular location in the texts. For instance, three possible locations are considered: "Begin", "Middle" and "End". The first sentence of a text is considered to be located at "Begin", the last sentence is said to be located at "End", and all other sentences are located at "Middle" in the text. The value of $P(N_y|SS_i,$Location) is obtained diving the number of summary sentences generating N_y cluster sentences at *Location* by the total number of summary sentences in the same *Location*.

It is important to say that not all Cluster Sentences are generated by the factors mentioned above. For this reason, we introduce $P(N_z|$None) which is the probability of N_z sentences being generated without the influence of any of the factors mentioned above or by some still unknown factor. The value of $P(N_z|$None) is obtained dividing the number of cluster sentences associated to none of the MDS factors, by the total number of cluster sentences. The union of all of the probabilities described above formulate P(C|S). This is shown in (3).

$$P(C \mid S) = \prod_{j} \prod_{i} P(N_x \mid SS_i, F_i) \times P(F_i, SS_i) \times P(N_y \mid SS_i, Location) \times P(N_z \mid None) \quad (3)$$

3. A brief example

Let's consider an hypothetical corpus composed only by a two-text cluster on the topic of the earthquake of Japan. These texts and their CST relations are shown in Figure 2.

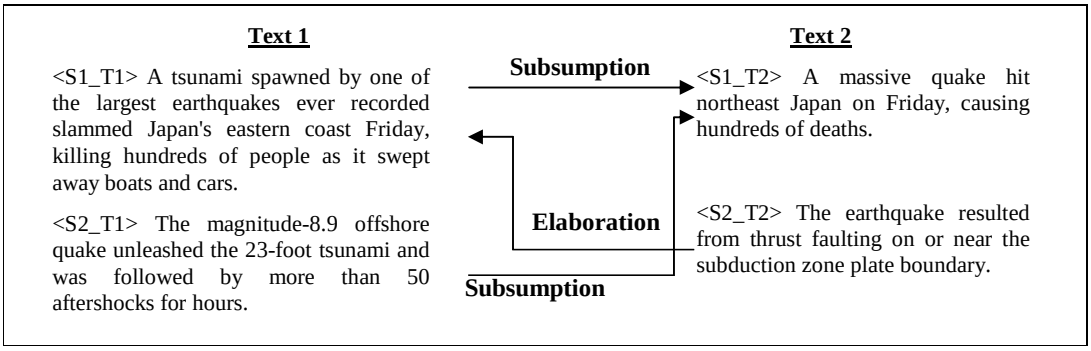

Figure 2. Example of Cluster of texts and CST Relations

Lets also consider that the correspondent extractive summary for this one-cluster corpus is a summary composed of two sentences only: <S1_T2> and <S2_T2>. According to this, we learn the probabilities, which are the model parameters:

$P(2 \mid SSi, \text{Redundancy}) = 1/2 = 0.5$

$P(1 \mid SSi, \text{Complementarity}) = 1/2 = 0.5$

$P(0 \mid SSi, \text{Contradiction}) = 2/2 = 1$

$P(1 \mid SSi, \text{Begin}) = 2/2 = 1$

$P(1 \mid SSi, \text{End}) = 1/2 = 0.5$

$P(\text{Complementarity} \mid SSi) = 1/2 = 0.5, etc...$

Figure 3. Probability values for example in Figure 2

It is important to notice that some probabilities will obtain value 0, since they may represent patterns that don't occur in the corpus. In this case, we may smooth those values by assigning a very small value close to 0, for example 0. 00001.

Once we have these parameters trained, we do the decoding process. In this stage we generate all possible extractive summaries for a given cluster and instantiate into the P(S|C) formula. For example, let's consider two candidate summaries each containing 1 sentence: <S1_T1> and <S2_T1> respectively:

$P(\text{Summary1} \mid C) =$ $P(1 \mid S1\_T1, \text{Redundancy}) \times P(\text{Redundancy} \mid S1\_T1)$ $\times P(1 \mid S1\_T1, \text{Complementarity}) \times P(\text{Complementarity} \mid S2\_T1)$ $\times P(0 \mid S1\_T1, \text{Contradiction}) \times P(\text{Contradiction} \mid S1\_T1)$ $\times P(1 \mid S1\_T1, \text{Begin}) \times P(0 \mid S1\_T1, \text{Middle}) \times P(1 \mid S2\_T1, \text{End})$ $\times P(0 \mid \text{None})$	$P(\text{Summary2} \mid C) =$ $P(1 \mid S2\_T1, \text{Redundancy}) \times P(\text{Redundancy} \mid S2\_T1)$ $\times P(0 \mid S2\_T1, \text{Complementarity}) \times P(\text{Complementarity} \mid S2\_T1)$ $\times P(0 \mid S2\_T1, \text{Contradiction}) \times P(\text{Contradiction} \mid S2\_T1)$ $\times P(1 \mid S2\_T1, \text{Begin}) \times P(0 \mid S2\_T1, \text{Middle}) \times P(0 \mid S2\_T1, \text{End})$ $\times P(0 \mid None)$

Figure 4. Values for P(Summary1|C) and P(Summary2|C)

After doing all the calculations, we obtain a value of 12.5×10^{-11} for P(Summary1|C) and a value of 12.5×10^{-12} for P(Summary2|C). In this example, Summary1 outperforms Summary2.

4. Final Remarks

In this paper we have presented a generative approach for MDS using the Noisy-Channel model, semantic-discursive information provided by CST and some other superficial features such as sentence position. One of the main contributions of this theoretical model is that it allows exploring the process of summary generation by analyzing different MDS factors. In future works we intend to turn this initial idea into a more sophisticated model that includes rhetorical information as another way to explore information generation. We also plan to investigate the most adequate Language model for P(S). Finally, we will explore heuristics for the decoding process, since for every possible summary, the probability P(S|C) has to be calculated and, depending on the database size, this can be a very expensive task.

Acknowledgements

The authors are thankful to FAPESP, CAPES and CNPq for their support.

References

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In the Proceedings of the 14th National Conference on Artificial Intelligence (AAAI), pp. 622-628. American Association for Artificial Intelligence.

Ng., A. and Jordan, M. (2001). On Discriminative vs. Generative classifiers: A comparison of logistic regression and Naive Bayes. *Neural Information Processing Systems*.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal.*