A Generative Approach for Multi-Document Summarization using the Noisy Channel Model

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. This task has become very important, since it assists the information processing in days where the amount of information is growing considerably. In this paper, we propose a statistical generative approach for multi-document summarization. In particular, we formulate the multi-document summarization task using a Noisy-Channel model. This approach is novel for multi-document summarization and it explores the process of summarization through the analysis of factors, such as redundancy, complementarity and contradiction. In this work, we model these factors using the Cross-document Structure Theory.*

1. Introduction

Multi-Document Summarization (MDS) is the process of building a summary from a group of texts that have similar content (Mani, 2001; McKeown and Radev, 1995; Radev and McKeown, 1998). This task shows to be a very useful resource in a scenario where the information available in the internet is growing exponentially, and new technologies that deal with this information are emerging.

MDS appeared as an extension of Single-Document Summarization, which its main purpose is to build a summary from a single document. However, MDS deals with different phenomena such as redundant, complementary and contradictory information. These phenomena appear because MDS treats various texts with diverse writing styles and a common topic. For example, if various newspapers are reporting about the earthquake in Japan, many of them may overlap information about the epicenter while some others may contradict on the number of deaths.

According to Mani and Maybury (1999), there are two main approaches for Automatic Summarization (single and multi-document summarization): superficial and deep approaches. Superficial approach involves statistical/empirical methods (e.g. word frequency methods, position of phrases/sentences in a text, etc.) that use little linguistic knowledge. These methods are said to produce summaries of low quality, but usually they have a low processing cost and, and they tend to be independent of the language. Investigations in this line include works based on simple word frequency methods, such as Luhn (1958), Edmundson (1969), Pardo et al. (2003), Radev et al.(2000), and more complex methods that use machine learning techniques, such as Kupiec et al. (1995), Mani and Bloedorn (1998), Chuang and Yang (2000), Larroca et al. (2002). On the other hand, deep approach involves methods that use sophisticated linguistic knowledge like grammars, semantic and/or discourse information. Some relevant works in this line include ontology-based investigations such as Afantenos et al. (2004), Afantenos (2007) and Henning et al. (2008). Other important works focus on semantic relations among documents, which represent similarities, differences and complements among different sources of information. Important investigations that use these type of relations are: Mani and Bloedorn (1997), Radev and McKeown (1998) and the works based on the recently explored Cross-

Anais do III Workshop "A RST e os Estudos do Texto", páginas 75–87, Cuiabá, MT, Brasil, Outubro 24–26, 2011. ©2011 Sociedade Brasileira de Computação

Document Structure Theory (CST) (Radev, 2000); such as Zhang et al. (2002), Otterbacher et al. (2002), Jorge and Pardo (2010) and Maziero et al. (2010). CST proposes a set of 24 semantic-discursive relations that represent the factors involved in MDS.

Another way of classifying MDS methods is into Generative or Discriminative approaches (Ng and Jordan, 2001). In general terms, in a generative approach, a model for some particular task is learned from the joint probability P(x,y) of inputs *x* and output labels *y*. Predictions are made by using Bayes Rule to calculate p(x|y) and then choosing the most likely *y*. This approach allows generating the observable data and permits the exploration of the generation process of these data. On the other hand, in the discriminative approach, the probability p(y|x) is modeled directly by mapping each x to the correspondent y. This is a common classification problem in which we want to determine the class of an element. A generative approach for AS provides methods that explore the summary building process, generate various possible summaries and search for the most likely one given a text or cluster of texts. On the other hand, a discriminative approach for AS provides methods that directly map a text (or group of texts) to the correspondent summary.

In the discriminative perspective, works like Schilder and Kondadadi (2008) and Aker et al. (2010) explore machine learning techniques with different textual features such as sentence position, word-frequency, semantic information, etc. In the generative perspective, there are a few works for MDS such as Haghighi and Vanderwende (2009) and Daumé III and Marcu (2006). Haghighi and Vanderwende explore generative probabilistic models based on the divergence of word distributions between summaries and document sets, while Daumé III and Marcu explore a Bayesian Model for query focused MDS.

The few investigations for generative and discriminative approaches may reveal a difficulty in exploring MDS through this approach. The reason for this may rely in the fact that exploring the factors within MDS is a high-cost task that requires various resources (e.g. annotated corpus, semantic-discursive parsers) providing deep information about these factors.

In this work we explore the generative process of MDS through the analysis of factors like redundancy, contradiction and complementarity and the formulation of MDS using the Noisy-Channel model. We explore the content selection task for summary generation through this model, by using CST semantic relations to represent the MDS factors mentioned above. This novel approach yields a theoretical generative learning model for MDS, which may improve its complexity by including more factors in the process. We expect this to be a matter of interest for future investigations.

This paper follows with a review of previous works and a novel proposal for MDS, particularly in sections 2 and 3 respectively.

2. Previous Work

Different areas in Natural Language Processing have projected their goals through a generative perspective by using the Noisy-Channel model; for example, Machine Translation (Brown et al., 1993), Question Answering (Echihaby and Marcu, 2003), Sentence Compression (Knight and Marcu, 2002) and Automatic Summarization (Daumé III and Marcu, 2002).

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.

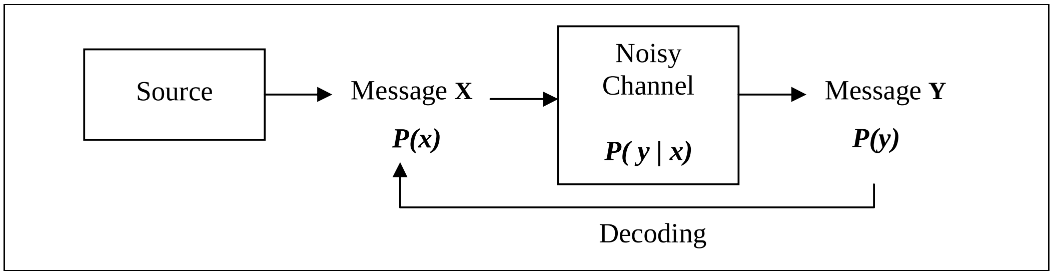

Figure 1: Noisy Channel Model

The noisy-channel works like this: a source produces an original message which passes through a channel where some noise occurs, 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 last stage is the focus of many applications that use this framework. This whole process is formulated through the Bayes rule:

$$P(x \mid y) = \frac{P(y \mid x) \times P(x)}{P(y)} \qquad (1)$$

In this formula, P(x|y) is the probability of decoding **x** given **y**. In this probability, we want to find the value of **x** that maximizes the result of (1). For this, we have to do the inverse calculus first: P(y|x), which corresponds to the probability of occurring message **y** given the original message **x**. Notice that P(y|x) is influenced by the noise introduced in the channel. At this point, the generative process for any task consists in determining which possible noises could be introduced in **x** so that **y** is generated. It is also important to point out that P(y) is going to be a constant value when trying to find the x that maximizes the result of P(x|y) for a given y. This is because P(y) is an observed value for all the data set. Moreover, the probability P(x) may be given by an adequate language model which varies depending on the task where the noisy-channel is applied.

As mentioned before, various works have used the Noisy-Channel model for different tasks. In the line of AS one of the most important investigations using Noisy-Channel is the proposal of Knight and Marcu (2002). In this work, the authors proposed a probabilistic model for sentence compression that preserves the grammaticality and relevance of information. For this aim, the sentence compression was modeled through a Noisy-Channel framework where the original message is a compressed sentence that passes through the noisy channel and produces a bigger sentence. Sentences are represented through syntactic trees produced by the Collins' Parser (Collins, 1997), and probabilities are computed over the syntactic components of the tree structure. The probabilities express the chance of a component to expand in a bigger component, which corresponds to the structure of the bigger sentence. The method was trained over the Ziff Davis corpus, which is a parallel corpus of documents and abstracts. In this corpus, the authors identified pairs of sentences corresponding to the abstract and the respective document. This paring represents the original sentence and the compressed version. The probabilities P(y|x) are computed during training. Some of the sentences of the corpus where separated for testing. At this point, all possible compressions for each of these sentences were generated, and then, the most likely one was chosen according to the model obtained at the training phase. Results in this work showed that this method produces grammatical and relevant sentences, in contrast with the sentences produced by baseline algorithms.

Another interesting work using Noisy-Channel is Daumé III and Marcu (2002), on document compression. The idea of the authors was to model the compression of a single document in a similar way to the work of Knight and Marcu, but also using rhetorical information. For this, the authors combined in a unique structure the rhetorical structure components of a text and the syntactic structure components of the elementary discourse units (EDU) of the text. They first constructed the rhetorical tree for each text and, then, for each EDU, a syntactical parsing was applied. The idea was to model the expansion of a summary into a bigger text by expanding the discourse constituents in the rhetorical tree of the summary. The authors used a journalistic corpus in which texts were paired with summaries and their correspondent rhetorical trees, extended with the syntactic structure of each EDU. Their system was evaluated with various baseline systems: human compression, random drop-word system and Concat (compresses all sentences of the text with the system of Knight and Marcu, and then concatenates them). Results showed that using discourse information leaded to more grammatical and coherent results when compared to automatic baseline systems.

Steinberger et al. (2010), in a similar way to the works described above, proposed the use of a Noisy-Channel model with a Phrase-Based approach (Brown et al., 1993) for sentence reconstruction. The idea was to reconstruct sentences of a summary turning them into bigger sentences by introducing more elements. For this, sentences were modeled as a set of words without considering stopwords. Similar to the other systems studied in this section, the authors used a parallel journalistic corpus. Their system was evaluated against the systems participating in TAC 2009, showing a good performance.

Other generative approaches like Haghighi and Vanderwende (2009) and Daumé III and Marcu (2006) do not use a Noisy-Channel model but rely in simple Bayesian statistics using information gain measures such as Kullback–Leibler divergence[1].

In the next section, a Noisy-Channel model for multi-document summarization is described.

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

Following the idea of other areas such as Statistical Machine Translation, Sentence Compression and Single-document Compression, we use the Noisy-Channel model for MDS.

When instantiating MDS in the Noisy-Channel framework, we adequate it into the three parts of the model: The source, the channel and the decoder. Initially, we assume that our source will produce a multi-document summary, which is a short text containing the most relevant information from a group of texts on a same topic. The probability for this summary is expressed by *P(S)* and it represents the chance of the summary to be a good summary. In this probability, many factors may be considered such as grammaticality, coherence, cohesion and relevance of information. All these features are expressed through a language model for summaries. This language model could be given, for example, by any summary evaluation metric such as Bleu (Papineni et al., 2002) or Rouge (Lin and Hovy, 2003).

In the next stage, the summary passes through the noisy channel, where some noise is introduced, and so, a cluster (group) of texts on the same topic is produced. This is expressed through P(C|S), which is the probability of producing an expanded cluster of texts from a summary. Finally, in the decoding stage the goal is to combine P(S) and P(C|S) to obtain P(S|C), which is the formulation of the noisy channel model for MDS through the Bayes Rule (2). Here, a set of possible summaries will instantiate the Bayes Rule in order to obtain the best summary, taking into account the probability P(C|S) calculated in the channel model and the language model P(S). The instantiated Noisy-Channel model for MDS is shown in Figure 2.

[1] Kullback–Leibler divergence is a non-symetric measure for calculating the difference between two distributions

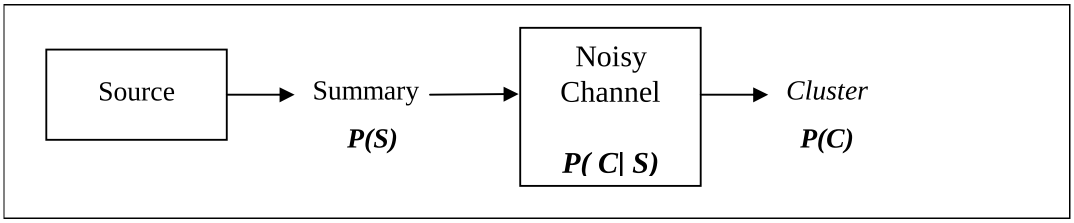

Figure 2: Noisy-Channel for MDS

$$P(S \mid C) = \frac{P(C \mid S) \times P(S)}{P(C)} \qquad (2)$$

In this work we will concentrate on the exploration of the channel model, P(C|S), in order to formalize a theoretical generative model for the MDS task. 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 for all the texts in the cluster. We detail the channel model in the next subsection.

3.1 The Channel Model

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, let's imagine the next sentence as a small hypothetical summary:

1) A massive 8.9-magnitude quake hit northeast Japan on Friday, causing dozens of deaths.

This sentence could generate other complementary information in any of the original texts such as,

2) The earthquake on March 11, 2011, resulted from thrust faulting on or near the subduction zone plate boundary between the Pacific and North America plates.

and/or redundant information,

3) The magnitude 9.0 Tohoku earthquake on March 11, 2011, occurred near the northeast coast of Honshu, Japan.

As it can be observed from the examples, the generative process occurs when the information of the sentences of a summary is expanded, by the insertion of new sentences that have redundant, complementary and contradictory information. This information can be given by any model representing these MDS factors; in particular, CST provides this type of information, through semantic relations across sentences of multiple documents. For example, according to CST, the second sentence of the example above has an "Elaboration" relation with the first sentence, since the information of the second elaborates the content of the first one; in other words, complementary information for the first sentence is given. Previous works on the topic of MDS have shown that MDS factors can be modeled through CST (Jorge and Pardo, 2009; 2010; Maziero et al., 2010). For example, complementary information can be modeled through relations: *Elaboration, Historical background* and *Follow up*; redundant information can be modeled through relations: *Identity, Equivalence, Subsumption, Overlap* and *Summary*; and contradiction can be modeled through *Contradiction* relation. In Figure 3, a small example shows how these relations occur in a group of sentences from different texts from the same cluster.

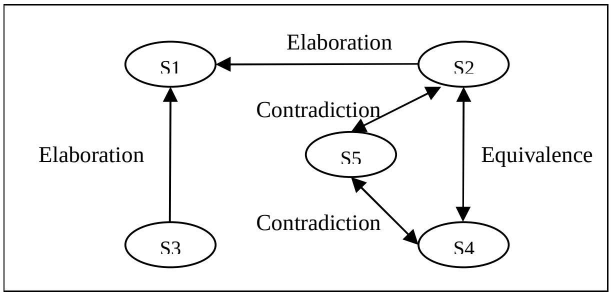

Figure 3: Example of CST relations between sentences

In this example we can observe that sentence 5 (S5) contradicts the information of sentences 2 and 4 (S2 and S4, respectively) which are similar in their content; sentence 2 and sentence 3 (S3) elaborate the information in sentence 1 (S1).

In the generative process, each sentence of the summary is potentially expansible, since it may produce extra sentences through any of the factors described above. We can formalize this generative process by establishing three initial conditions:

- A summary is a set of sentences $SS=\{ SS_1, \ldots, SS_n\}$
- 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: $CS= \{CS_1, \ldots, CS_m\}$. These sentences are influenced by the MDS phenomena factors which should be made explicit (e.g. sentences could be annotated with CST relations).
- A set of MDS phenomena factors is given, $F=\{F_1, \ldots, F_z\}$

Initially, the factors used in the generative process can be represented by the CST relations, but this is not mandatory since other forms of modeling these factors may be used. Once we have established these points, we can formalize the generative process with the algorithm shown in Figure 4.

```
For each sentence $SS_i$ of the summary
    For each MDS phenomena factor $F_i$
        If $F_i$ applies to $SS_i$ then
            create N sentences for $SS_i$ that represent $F_i$
            include the N sentences in the CS set
```

Figure 4: Algorithm for noise generation in MDS

To build this generative model, we consider having a parallel multi-document corpus containing clusters of texts annotated with MDS factors (e.g., sentences annotated with CST relations) and their correspondent extractive summaries (summaries built with sentences extracted from the original texts). Once we have the corpus available, P(C|S) is calculated by multiplying probabilities describing the chance of a sentence SS_i to generate a quantity N of sentences through factor F_j. This is formulated in (3)

$$P(C \mid S) = \prod_{j=1} \prod_{i=1} P(N \mid SSi, Fj) \tag{3}$$

The value of *P(N|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. For example, let's suppose there is a corpus containing 6 summary sentences, 2 of them generating 3 redundant cluster sentences; in this case, *P(3|SSi, Complementarity)* is calculated by 2/6 which results in 0.33. In other words, the chance of a summary sentence to generate 3 complementary sentences is 0.33. The probabilities for the three MDS factors are expressed in (4), (5) and (6).

$$P(N_x \mid SS_i, \text{Redundancy}) \tag{4}$$

$$P(N_y \mid SS_i, \text{Complementarity}) \tag{5}$$

$$P(N_z \mid SS_i, \text{Contradiction}) \tag{6}$$

A probability *P(F$_j$|SS$_i$)* is associated to each probability template, 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$_i$. For example, let's suppose we have a corpus containing a certain number of Cluster Sentences, which 4 of them are associated to the Complementarity factor out of 6 generated by summary sentences associated to any other MDS factor, then *P(Complementarity|SS$_i$)* is 4/6. In other words, the chance of a summary sentence to generate sentences by Complementarity factor is 0.66 The union of the probabilities *P(N|SS$_i$, F$_j$)* and *P(F$_j$|SS$_i$)* is expressed in (7), (8) and (9).

$$P(N_x \mid SS_i, \text{Redundancy}) \times P(\text{Redundancy} \mid SS_i) \tag{7}$$

$$P(N_y \mid SS_i, \text{Complementarity}) \times P(\text{Complementarity} \mid SS_i) \tag{8}$$

$$P(N_z \mid SS_i, \text{Contradiction}) \times P(\text{Contradiction} \mid SS_i) \tag{9}$$

Another generative factor considered in our model is the location of the cluster sentences. For this, we associate a probability *P(N|SS$_{i,}$ Location),* which expresses the chance of SS$_i$ generating a number N 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|SS$_{i,}$Location)* is obtained diving the number of summary sentences generating N cluster sentences at *Location*, by the total number of summary sentences in the same *Location*. For example, let's consider a corpus containing a certain number of Cluster Sentences, 3 of them generated by any F$_i$ and 2 out of those 3 are generated at the beginning of a text, then, the value of P(2|SSi,Begin) is 2/3 or 0.66. In other words, the chance of a summary sentence to generate 2 cluster sentences at the beginning of the texts is 0.66.

It is important to say that not all Cluster Sentences are generated by the factors mentioned above. For this reason, we introduce *P(N|None)* which is the probability of N sentences being generated without the influence of any of the factors mentioned above or by some still unknown factor. The value of *P(N|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 probability templates described above formulates P(C|S). This is shown in (10).

$$
\begin{aligned}
P(C \mid S) = \prod_{i=1}^{M} & P(N_x \mid SS_i, \text{Redundancy}) \times P(\text{Redundancy} \mid SS_i) \\
& \times P(N_y \mid SS_i, \text{Complementarity}) \times P(\text{Complementarity} \mid SS_i) \\
& \times P(N_z \mid SS_i, \text{Contradiction}) \times P(\text{Contradiction} \mid SS_i) \\
& \times P(N_u \mid SS_i, \text{Begin}) \times P(N_v \mid SS_i, \text{Middle}) \times P(N_w \mid SS_i, \text{End}) \times P(N \mid None)
\end{aligned} \tag{10}
$$

The calculus of (10) is done for one single Cluster Summary, in a real scenario, the counts must be done for every possible summary and the one obtaining a higher value will be the most suitable summary for the given cluster, this is the decoding task.

3.2. A brief example

Let's consider an example of a two-text cluster which is shown in Figure 5.

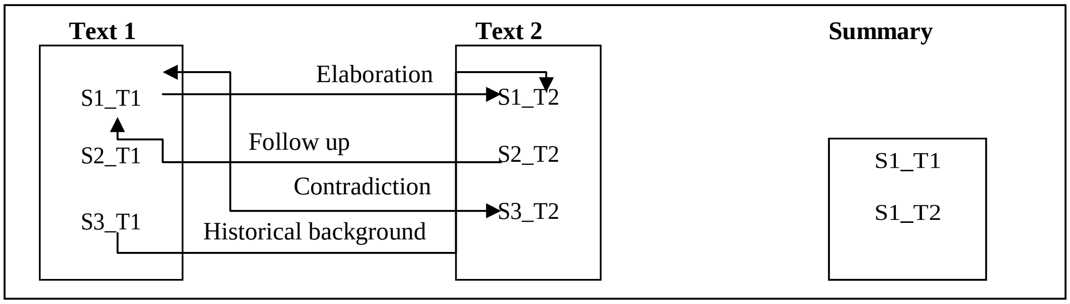

Figure 5: Example of a multi-document cluster with correspondent summary

For the training of the channel model (described in the previous sub-section), we assume that the example in Figure 5 represents a hypothetical parallel corpus annotated with CST relations. It can be observed that the parallel summary is an extract that contains two sentences of the cluster: sentence 1 of Text 1(S1_T1) and sentence 1 of Text 2 (S1_T2).

Having this corpus, we learn the probabilities, which are the model parameters. As an example of this training, we show in Figure 6 some of the probability values that were extracted from the example in Figure 5.

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

$$\text{P}(0 \mid \text{SSi, Complementarity}) = 0/2 = 0.0000001$$

$$\text{P}(0 \mid \text{SSi, Contradiction}) = 1/2 = 0.5$$

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

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

$$\text{PP(Complementarity} \mid \text{SSi)} = 3/5 = 0.6$$

$$\text{P(Contradiction} \mid \text{SSi)} = 1/5 = 0.2$$

$$\text{P(Redundancy} \mid \text{SSi)} = 0/5 = 0.0000001 \text{ ...etc}$$

Figure 6: Probability values for example in Figure 5

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 smooth those values by assigning a very small value close to 0, for example 0. 0000001. It can also be used more sophisticated smoothing techniques, but this will be explored in future works.

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. In this case, P(S|C) will be the same as P(C|S) since we are considering P(S) uniform and P(C) constant. Notice that different summaries will produce different probability values. For example, let’s consider the candidate summaries 1 and 2 in Figure 7, and their correspondent values for P(Summary 1|C) and P(Summary 2|C) in Figure 8.

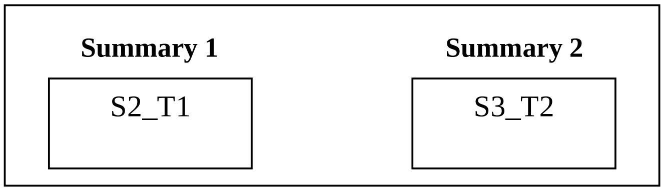

Figure 7: Candidate summaries for example in Figure 5

$$\begin{aligned}
&\text{P(Summary1} \mid \text{C)} = \text{P(C} \mid \text{Summary1)} = \\
&\text{P}(0 \mid \text{S2\_T1, Complementarity}) \times \text{P}(0 \mid \text{S2\_T1, Contradiction}) \\
&\times \text{P}(0 \mid \text{S2\_T1, Redundancy}) \times \text{P(Complementarity} \mid \text{S2\_T1)} \\
&\times \text{P(Contradiction} \mid \text{S2\_T1)} \times \text{P(Redundancy, S2\_T1)} \\
&\times \text{P}(0 \mid \text{S2\_T1, Begin}) \times \text{P}(0 \mid \text{S2\_T1, Middle}) \times \text{P}(0 \mid \text{S2\_T1, End}) \\
&\times \text{P}(6 \mid \text{None})
\end{aligned}$$

$$
\begin{aligned}
&\mathrm{P(Summary2} \mid \mathrm{C)} = \mathrm{P(C} \mid \mathrm{Summary2)} = \\
&\mathrm{P(0} \mid \mathrm{S3\_T2, Complementarity)} \times \mathrm{P(1} \mid \mathrm{S3\_T2, Contradiction)} \\
&\times \mathrm{P(0} \mid \mathrm{S3\_T2, Redundancy)} \times \mathrm{P(Complementarity} \mid \mathrm{S3\_T2)} \\
&\times \mathrm{P(Contradiction} \mid \mathrm{S3\_T2)} \times \mathrm{P(Redundancy} \mid \mathrm{S3\_T2)} \\
&\times \mathrm{P(1)} \mid \mathrm{S3\_T2, Begin)} \times (\mathrm{P(0} \mid \mathrm{S3\_T2, Middle)} \times \mathrm{P(0} \mid \mathrm{S3\_T2, End)} \\
&\times \mathrm{P(5} \mid \mathrm{None)}
\end{aligned}
$$

Figure 8: Values for P(Summary 1|C) and P(Summary 2|C)

Summary 1 is composed by only one sentence of the cluster, S2_T1, which generates 0 complementary sentences (sentences that complement the information of S2_T1), 0 redundant sentences and 0 contradictory sentences. In consequence, it generates 0 sentences at the "Beginning", "Middle" and "End" locations. And there are 6 sentences generated without any influence of the MDS factors. On the other hand, Summary 2 is composed by sentence S3_T2 which generates 1 contradictory sentence, 0 complementary sentences and 0 redundant sentences. It generates 1 sentence at the "Beginning", 0 sentences at the "Middle", 0 sentences at the "End" and 5 sentences are generated without any influence of the MDS factors.

After doing all the calculations, we obtain a value of 3×10^{-37} for P(Summary1|C) and a value of 15×10^{-31} for P(Summary2|C). In this example, Summary 2 outperforms Summary 1, which means Summary 2 is the best summary according to our model. It is important to notice that the probability values are very small; this may be because this hypothetical corpus is composed by a small number of sentences and data will tend to be sparse. This may be a problem since small values may be rounded to zero. In a real scenario, these values will tend not to be too small, since there will be much more examples in the corpus and smoothing will be carried out.

4. Final Remarks

In this paper we have presented a Noisy-Channel model for multi-document summarization by considering the MDS phenomena factors as parameters for the generative approach. In particular, it was illustrated a modeling of these factors using semantic-discursive information provided by CST and some other superficial features such as sentence position. The main contribution of this work is a theoretical model for generative MDS using the Noisy-Channel framework. One of the main advantages of this model is that it allows exploring the process of summary generation by analyzing different factors which may be represented by CST or any other semantic-discursive model. Another advantage of this model is that it allows searching for the most likely summary by exploring the factors that influence in the informativeness of these summaries. This is the first generative approach through Noisy Channel. In future works we intend to turn this initial idea into a more sophisticated model that includes rhetorical information and more text surface characteristics. We also plan to investigate the most adequate Language model for P(S), so we can look for the most likely summary not only in terms of informativeness, but also in terms of grammar, coherence, cohesion, etc.

Besides the advantages of this model, some limitations have to be pointed out. For example, in an eventual empirical evaluation of the model, the training and testing will be done on extract summaries only, since with extracts it is easier to explore the generation of sentences through CST relations. Another limitation is that the generation of sentences does not distinguish among texts, in other words, a set of sentences is generated without distinguishing to which text the sentence belongs to. Finally, another disadvantage is the cost of the decoding process, since for every summary has to be calculated P(S|C) and depending on the database size this can be a very expensive task.

Acknowledgements

The authors thank FAPESP, CNPq and CAPES for their support in this work.

References

Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*, pp. 410-419.

Afantenos, S.D. (2007). Reflections on the Task of Content Determination in the Context of Multi-Document Summarization of Evolving Events. In Recent *Advances on Natural Language Processing* 2007. Borovets/Bulgaria.

Aker, A.; Cohn, T.; Gaizauskas, R.(2010). Multi-document summarization using A* search and discriminative training. In the Proceedings of the *Conference on Empirical Methods in Natural Language Processing of ACL*. Stroudsburg/USA.

Banko, M.; Mittal, V.; Witbrock, M. (2000). Headline generation based on statistical translation, In the Proceedings of the *38th Annual Meeting of the ACL*, pp. 318–325. Hong Kong.

Brown, P. E.; Pietra, S. A. D.; Pietra, V. J. D.; Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 16, N. 2, pp. 79-85.

Collins, M. (1997). Three generative lexicalized models for statistical parsing. In the *Proceedings of the 35th Annual Meeting of the ACL*, pp. 16–23. Madrid/Spain.

Daumé III, H. and Marcu, D. (2002). A noisy-channel model for document compression. In the *Proceedings of the Conference of the Association for Computational Linguistics*, pp. 449-456. Philadelphia/USA.

Daumé III, H. and Marcu, D. (2006). Bayesian query-focused summarization. In the *Proceedings 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* - ACL. Sydney/Australia

Echihabi, A. and Marcu , D. (2003). A noisy-channel approach for question answering, In the *Proceedings of Association for Computational Linguistics*-ACL. Sapporo/Japan

Edmundson, H. P. (1969). New Methods in automatic extracting. *Journal of the ACM*, Vol. 16, pp. 264-285

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In the *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*-NAACL. pp.362-370. Boulder/Colorado

Hennig L.; Umbrath W.; Wetzker R. (2008). An Ontology-Based Approach to Text Summarization. In the *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology* - IEEE/WIC/ACM, pp.291-294. Sydney/Australia

Jorge, M.L.C. and Pardo, T.A.S. (2009). Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-8. September 8-10, São Carlos/SP, Brazil.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. July 16, Uppsala/Sweden

Jorge, M.L.C; Agostini, V.; Pardo, T.A.S. (2011). Multi-Document Summarization using Complex and Rich Features. In the *Proceedings of the XXXI Conference of the Brazilian Computing Society-CSBC*. Natal/Brazil.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, V. 139, N.1, pp. 91-107

Kupiec, J.; Pedersen, J.; Chen, F. (1995). A trainable document summarizer. In the *Proceedings of the 18th ACMSIGIR Conference on Research & Development in Information Retrieval*, pp. 68-73. Washington/USA.

Larocca Neto, J.; Freitas, A.A; Kaestner, A.A.C. (2002). Automatic Text Summarization using a Machine Learning approach. In the *Proceedings of the 16th BrazilianSymposium on Artificial Intelligence*, pp. 205-215. Porto de Galinhas/Recife.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton/Canada

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development,* Vol. 2, pp. 159-165.

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.

Mani, I. and Bloedorn, E. (1998). Machine Learning of Generic and User-Focused Summarization. In *Proceedings of the Fifteenth National Conference on* AI (AAAI-98), pp.821-826. Madison/ Wi USA

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science*- NLPCS, pp.60-69. June 8-12, Funchal/Madeira, Portugal.

McKeown, K. and Radev, D.R. (1995). Generating summaries of multiple news articles. In the *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-82, Washington/USA.

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

Otterbacher, J.C.; Radev, D.R.; Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*, pp 27-36. Philadelphia/USA.

Papineni, K.;Roukos, S.; Ward, T.;Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics-ACL*, pp.311-318. Philadelphia/USA.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR *(Lecture Notes in Artificial Intelligence 2721)*, pp. 210-218 June 26-27, . Faro/Portugal.

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.

Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, N. 3, pp. 469-500.

Radev, D.R.; Jing, H.; Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In the *Proceedings of the ANLP/NAACL Workshop*, pp. 21-29. Washington/USA.

Schilder, F. and Kondadadi R. (2008). FastSum: fast and accurate query-based multi-document summarization. In the Proceedings of the 46th Annual Meeting of the ACL. Stroudsburg/USA

Steinberger,J.; Turchi, M.; Kabadjov,M.; Cristianini,N.; Steinberger R. (2010). Wrapping up a Summary: from Representation to Generation. In the *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 382-386. Uppsala/Sweden

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. 2002. Towards CST-Enhanced Sumarization. In the *Proceedings of the 18th National Conference on Artificial Intelligence*. Edmonton/Canadá.