Revista de Estudos da Linguagem, v. 26, n. 1, p. 45-71, 2018

Exploring content selection strategies for Multilingual Multi-Document Summarization based on the Universal Network Language (UNL)

Investigando estratégias de seleção de conteúdo para a Sumarização Multi-Documento Multilíngue com base na Universal Network Language (UNL)

Matheus Rigobelo Chaud
Universidade de São Carlos, São Carlos, São Paulo / Brasil
matheus_chaud@yahoo.com.br

Ariani Di Felippo
Universidade Federal de São Carlos, São Carlos, São Paulo / Brasil
arianidf@gmail.com

Abstract: Multilingual Multi-Document Summarization aims at ranking the sentences of a cluster with (at least) 2 news texts (1 in the user's language and 1 in a foreign language), and select the top-ranked sentences for a summary in the user's language. We explored three concept-based statistics and one superficial strategy for sentence ranking. We used a bilingual corpus (Brazilian Portuguese-English) encoded in UNL (*Universal Network Language*) with source and summary sentences aligned based on content overlap. Our experiment shows that "concept frequency normalized by the number of concepts in the sentence" is the measure that best ranks the sentences selected by humans. However, it does not outperform the superficial strategy based on the position of the sentences in the texts. This indicates that the most frequent concepts are not always contained in first sentences, usually selected by humans to build the summaries because they convey the main information of the collection.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.26.1.45-71

Keywords: content selection; concept; statistical measure; multilingual corpus; multi-document summarization.

Resumo: O objetivo da Sumarização Automática Multilíngue Multidocumento é ranquear as sentenças de uma coleção com ao menos duas notícias (1 na língua do usuário e 1 em língua estrangeira) e selecionar as mais bem pontuadas para compor um sumário na língua do usuário. Exploramos três estatísticas conceituais e uma estratégia superficial para criar um ranque das sentenças quanto à relevância. Para tanto, utilizamos um *corpus* bilíngue (português-inglês) anotado via UNL (Universal Network Language) e com textos-fonte e sumários alinhados em nível sentencial. A avaliação indica que a estatística denominada *frequência de conceitos normalizada pelo número de conceitos da sentença* é a que melhor reproduz o ranqueamento humano. Essa medida, entretanto, não supera a estratégia superficial baseada na *posição das sentenças*. Isso indica que os conceitos mais frequentes do *cluster* nem sempre estão contidos nas primeiras sentenças dos textos-fonte, usualmente selecionadas pelos humanos para compor os sumários porque veiculam a informação principal da coleção.

Palavras-chave: seleção de conteúdo; conceito; medida estatística; *corpus* multilíngue; sumarização multidocumento.

Received on: July 7th, 2016.

Approved on: January 13th, 2017.

1 Introduction

Even though a wide number of news agencies make information available on the web, it is very difficult to know what is happening in the World unless an event is tragic enough to catch the attention of the international media. According to Orasăn and Chiorean (2008), there are two main reasons for that. First, quite often the news is not in a language familiar to the reader. And second, even in the cases where the language does not constitute an impediment, the amount of information available is quite often so large that it is impossible to read everything published.

Thus, Natural Language Processing (NLP) applications that address the goal of treating multiple languages in different multi-document summarization tasks are relevant tools to deal with the huge and overloaded amount of information in multiple languages. One of these applications is the *cross-language summarization,* which is the production of a summary in a language Lx when the cluster (i.e., cluster of news texts on the same topic) is in a language Ly different from Lx (SARKAR, 2014).

Another application is called Multilingual Multi-Document Summarization (MMDS). In a broad sense, the definition of MMDS is: “If L is a set of natural languages, MMDS can be defined as a process that can accept a single document in one language $l \in$L or can accept a cluster of related documents in one language or in different languages selected from L to produce a summary in the same language as the input or in a language chosen from L by the user (SARKAR, 2014). In particular, when the input is a cluster of related documents coming from different languages sources, MMDS is a highly challenging NLP task, since it requires merging content in different languages as well as dealing with the classical multi-document issues, such as capturing the most relevant content, and maintaining summary coherence/cohesion by treating redundancy. MMDS approaches can be broadly categorized as *language independent multilingual* summarizationand *language dependent multilingual* summarization. The approaches of the first category do not use much semantic or language specific information. They can make only some minimal assumptions about the language (e.g., that the text can be split into sentences and sentences further into words) and perform equally well on different languages without linguistic knowledge. These approaches usually have low cost and are more robust, but they produce poor results. The approaches of the second category utilize language specific knowledge such as morphological, syntactic and/or semantic information, retrieved from lexical resources (e.g., wordnet lexical databases and thesauri) and parallel corpora. Language specific knowledge is necessary for machine translation of documents from one language to another.

Specifically, the few previous *language dependent* MMDSmethods usually consist of two steps: (i) translation of the foreign texts and (ii) summarization (ROARK; FISHER, 2005; EVANS *et al*., 2005; TOSTA *et al*., 2013). The first step is performed by some machine-translation (MT) engine, producing a monolingual multi-document cluster. Then, an extractive multi-document summarization method is used to build the summaries, which sometimes treats redundancy. As for the summarization step, the extractive methods are predominantly superficial, based on features such as *word frequency* and *sentence position*, which are robust and have low cost, but produce poor results (KUMAR, SALIM, 2012). Tosta (2014), whose results were recently published by Di-Felippo *et al*. (2016), has proposed the first MMDS methods exclusively based on lexical-conceptual knowledge. The methods use the frequency of the nominal concepts in the cluster to score and rank sentences in their original languages. If sentences in the foreign language are selected for the summary, they are automatically translated to the user's language. The experiments were performed using a corpus of 20 clusters, and show that conceptual knowledge improves the linguistic quality of extracts.

Given the promising results of Tosta (2014) and Di-Felippo *et al*. (2016), we have explored the potential of 3 concept-based measures to capture human content selection strategies in MMDS: (i) *CF* (*concept frequency*), (ii) *CF*IDF* (*concept frequency corrected by the inverted document frequency*), and (iii) *CF/No. of Cs in S* (*concept frequency normalized by the number of concepts in the sentence*). The experiment was performed using 3 clusters from the CM2News *corpus* (TOSTA, 2014), whose source sentences were manually annotated with UNL (*Universal Network Language*) (UCHIDA *et al*., 1999). To analyze the measures, we used manual alignment of the texts and human summaries at sentence level. Specifically, we calculated how many aligned source-sentences were covered by the top sentences of the ranks built from each measure and by a *sentence position* baseline. The experiment shows that measure (iii) produces the rank with the highest number of aligned sentences, having thus the best performance in capturing the human preferences. However, it did not outperform the *sentence position* baseline. This indicates that the sentences that convey the most important information in news texts are, indeed, in the initial positions, and also that they do not necessarily contain the most frequent concepts. This evidence, however, needs to be well explored due to our small corpus of work.

In Section 2, we detail researches that address the goal of treating multiple languages in different multi-document summarization tasks, especially those that rely on language specific knowledge. In Section 3, we describe the corpus that was used, focusing on the pre-processing step. In Section 4, we present the 3 concept-based measures that we investigated. In Section 5, we discuss our evaluation, which measures how the conceptual statistics are able to select the same source sentences as humans to compose an extract. Lastly, we provide final remarks and directions for further work in Section 6.

2 Related works

Evans and Klavans (2004) have developed a multilingual version of the Columbia Newsblaster as a testbed for cross-language multi-document summarization. The system collects, clusters, and summarizes news documents from sources all over the world daily. It crawls news sites in many different countries, written in different languages, extracts the news from HTML pages, uses a variety of methods to translate the documents for clustering and summarization, and produces an English summary for each cluster.

Sarkar and Bandyopadhyay (2005) presented the architecture of multilingual summarization system for Indian languages. Basically, the system has three major components: (i) several monolingual news clusters, (ii) a multilingual news clusters, and (iii) a news summarizer. The monolingual news cluster receives a news stream from multiple online newspapers in its respective language, and directs them into several output news streams by using events. Next, the multilingual news cluster matches and merges the news streams of the same event but in different languages in a cluster. The task for the multilingual cluster is to align the news clusters in the same topic, but in different languages. The system summarizes the news stories for each event by creating clusters of sentences and selecting the representatives from each cluster to form the final summary.

Roark and Fisher (2005) take as input a cluster of some machine-translated and original (written and spoken) texts. The method ranks all the source sentences based on superficial features, and sets a high preference for original English sentences. The features are different versions of the *tf-idf, log-likelihood ratio,* and *log-odds ratio* lexical measures, and *position,* which increase the weight of sentences near the beginning of texts. The method was trained on a set of 80 clusters with translations and original English texts using a machine-learning algorithm, but there is no detail about the evaluation. One problem with this method is that, considering machine-translated texts as input, the summaries might contain ungrammatical sentences, since MT is far from perfect.

Evans *et al*. (2005) present an approach that identifies similarities and differences across texts written in different languages for summarizing topically clustered texts from two sources, English and machine translated Arabic texts. Specifically, they take as input a cluster with machine-translated and original texts. They only rank translated sentences, using a combination of deep (i.e., importance-signaling words, high-content verbs, and dominant concepts) and superficial features. Besides the sentence position in the source texts, the other superficial feature for sentence extraction is *length,* which penalizes sentences that are shorter or longer than a threshold. The sentences selected from the rank are replaced with similar ones from the English texts. For evaluation, they used the DUC 2004 corpus, which contains 24 topics with Arabic-to-English machine translations and English texts, and 4 human summaries. Using ROUGE (LIN, 2004), the evaluation shows that the similarity-based approach outperforms a *first-sentence* baseline. This method, therefore, uses some semantic aspects of the input, an advance over Roark and Fisher (2005), although it is clear that relevant content that occurs exclusively in the preferred language is not selected to build the summary.

Wan *et al*. (2010) present a cross-language multi-document summarization approach that was evaluated on the manual translated version of the DUC[1] 2001 dataset. In this approach, each English document set is summarized to produce a Chinese summary. The approach performs three main steps: (i) prediction of the translation quality of each English sentence in the document set; (ii) selection of the English summary sentences based on the translation quality and informativeness, and (iii) translation of the generated English summary to form the final Chinese summary.

Tosta *et al*. (2013) also take as input a cluster with machine-translated and original texts. The authors have proposed two MMDS approaches based on superficial features: *word frequency* and *sentence position* methods. And both avoid redundancy applying the *word overlap* measure. If an ungrammatical translated-sentence is selected, it is replaced with a similar sentence from the original text. The methods were intrinsically evaluated according to the linguistic quality of the summaries using the criteria of DUC (DANG, 2005): grammaticality,

[1] Document Understanding Conference (http://duc.nist.gov/)

non-redundancy, referential clarity, focus, and structure/coherence. In the manual evaluation, the *sentence position* method had better results. Although the methods avoid the MT problems by applying a late-translation approach, the content selection still relies on flat features, which produce summaries with lower linguistic quality.

Tosta (2014) proposed 2 deep methods that take the source texts in their original language as input. Both methods use the frequency of the nominal concepts in the cluster to score the sentences, and avoid redundancy using *word overlap*. Given the rank, the CF (*concept frequency*) method selects the best-ranked sentences to compose the summary until the desired summary length is achieved. If a sentence happens to be in the foreign language, it is automatically translated to the user's language. The method was proposed under the assumption that the MT of the selected foreign sentences to the user's language minimizes the problems that are caused by full MT of the source texts in the summaries. The CFUL (*concept frequency + user language*) method selects the top-ranked sentences from the text written in the user's language to compose the summary, also avoiding redundancy. This approach relies on the assumption that a summary built exclusively with original sentences in the user's language reflects the most relevant content of the cluster, since the concepts that occur in the foreign text are also taken into account for sentence ranking. For evaluation, the authors used the CM2News corpus (TOSTA, 2014), which has 40 original news texts grouped by topic in 20 clusters. Each cluster contains 1 news text in English and 1 in (Brazilian) Portuguese, and 1 human summary in Portuguese. The goal was to produce extracts in Portuguese (user's language). The concepts were semiautomatically derived from Princeton WordNet. The evaluation using the DUC criteria showed that the conceptual knowledge improved the linguistic quality of the summaries, since both methods outperformed the *sentence position* baseline (TOSTA *et al*., 2013). It also showed that CFUL outperformed CF.

For summary evaluation, DUC was the main evaluation forum from 2001 until 2007. Nowadays, the Text Analysis Conference (TAC) provides a forum for assessment of different information access technologies including text summarization. Out of the past DUC and TAC editions, only a few have included multilingual text summarization tasks in the list of official tasks. Recently, TAC 2011 Summarization Track had a task on multilingual text summarization, which is called MultiLing.

Based on the cited works of the literature, we see that the most recent language-dependent approaches for determining important content in MMDS for English and Portuguese languages move towards a shallow semantic interpretation of summary language.

The lexical-conceptual knowledge has already been used in single summarization in order to achieve better content selection. Some methods start by indexing the words of a text to concepts of a domain-related taxonomy (i.e., hierarchy of concepts) and explore structural features of the taxonomy (e.g., *level*) to detect the main subtopics of the text (e.g., WU, LIU, 2003; HENNIG *et al*., 2008). Sentences or paragraphs that are "closer" to the subtopics are selected to compose the summary. Other approaches rely on the codification of the source text into UNL, and the application of different statistics for sentence scoring, picking the sentences with the highest score to build the summary (e.g., SORNLERTLAMVANICH *et al*., 2001; MANAGAIKARASI; GUNASUNDARI, 2012). Since the UNL is a formalism to express the propositional content of any sentence, Sornlertlamvanich *et al*. (2001), for example, remove redundant words from the selected sentences, such as modifiers, and combine sentences that cover the same concepts, producing abstracts. Pandian and Kalpana (2013) proposed an approach for summarizing documents from the tourism domain. The authors focused on the generation of summaries for different levels of users. Martins (2002) and Martins and Rino (2002) developed heuristic rules for single-document summarization at the intra-sentential level, which prune unnecessary binary relations from the UNL codification of a text.

The heavy reliance on language resources, such as WordNet and UNL formalism, is clearly a bottleneck for the aforementioned deep approaches, because success is constrained by the coverage of the resources and the sense granularity stored there. However, the use of conceptual knowledge generates better results than shallow approaches, at least in terms of linguistic quality.

Thus, this work focuses on: (i) exploring the potential of 3 concept-based statistics for determining important content in MMDS, (ii) using all kinds of concepts (not only nominal concepts), and (iii) evaluating the measures based on the alignments of source texts and human summaries at sentence level.

Next, we describe the UNL formalism and the pre-processing of the corpus.

3 The Corpus

3.1 The UNLization: conceptual annotation

Since we sought to investigate how content selection takes place in MMDS, we have selected the CM2News corpus (TOSTA, 2014; DI-FELIPPO, 2016). It has 40 original news texts (a total of 19,984 words) grouped by topic in 20 clusters. Each cluster is composed of 1 news text in English and 1 in (Brazilian) Portuguese, both on the same topic, and 1 human multilingual multi-document abstracts in BP. To produce the abstracts, the abstract-writers were instructed to produce summaries of length equal to 30% of the longest article in the cluster (i.e., 70% compression rate). The clusters cover different domains: world, politics, health, science, entertainment, and environment. Given the preliminary and exploratory nature of this work, we have selected only 3 clusters from CM2News, whose source texts and summaries have different sizes or lengths (in number of sentences and words) (Table 1).

TABLE 1 – Characteristics of the data collection (CHAUD, 2014)

Cluster	Topic/Domain	Reference	Document	Qt. sentences	Qt. words
C1	Attacks in London (World)	C1-PT	Source-text	17	518
		C1-EN	Source-text	36	788
		C1-Sum-ref	Reference summary	9	229
C2	Gay Kit (Politics)	C2-PT	Source-text	11	287
		C2-EN	Source-text	13	229
		C2-Sum-ref	Reference summary	4	84
C9	Earthquake in Missouri (World)	C9-PT	Source-text	25	511
		C9-EN	Source-text	33	660
		C9-Sum-ref	Reference summary	10	198
Total				158	3,504

The source texts and summaries received a layer of semantic annotation. In general, semantic annotation is additional information in a document that identifies or defines the semantics of a part of that document. In other words, we can say that semantic annotation is about attaching, for example, sense tags, names, attributes, comments, and descriptions to a document or to a selected part in a text.

For our annotation, we selected a specific formalism, called UNL (UCHIDA *et al*., 1999), which states that a deep semantic analysis for a natural language text requires two levels of semantics: lexical semantics and grammatical semantics. In particular, UNL expresses information conveyed by natural language (NL) sentences through binary relations between concepts. Thus, UNL is not different from the other formal languages devised to represent NL sentence meaning (MARTINS *et al*., 2002). The general syntax of the relations is *RL(UW1,UW2)*, where *RL* stands for a *Relation Label*, which signals the semantic relation, and *UWs* means *Universal Words*, which signal the related concepts. RLs are specified through mnemonics; they are three-letter symbols that signify the kind of semantic relationship that ties two UWs in a natural language utterance, for example, *agt* for *agent*, *mod* for *modifier*, or *obj* for *object*. *UWs* may be generic, such as *book*, or *John*, or complex, in which case they indicate meaning variations, for example, in *animal(icl>living thing)*, *icl* indicates a hyperonymic relation between *animal* and *livingthing*. *UWs* can also be annotated by attributes to provide further information on the circumstances under which they are used (e.g., tense and aspect). Those are signaled by *Attribute Labels* (*ALs*). According to Cardeñosa *et al*. (2008), the advantages of UNL are: (i) flexibility and neutrality, since it is a language to represent any content in any domain in any language, (ii) generality, since the set of *UWs* and *RLs* is sufficient to describe any kind of content expressed in NLs, and (iii) explicitness and clarity, which are univocal and machine-tractable.

Each cluster was manually annotated by 1 computational linguist in two-hours daily sessions, during 3 consecutive months, with the support of a tool called UNL Editor (ALANSARY *et al*., 2011). The UNL Editor is a visual tool designed with the intention of providing full semantic annotation, including the analysis of natural language texts and the generation of UNL documents. In particular, it provides a powerful visual interface for working with UNL data both in a textual and graphical mode with a friendly interface, creating an appropriate environment for navigating through the needed steps of providing the analysis. Most importantly, the UNL Editor's output offers the necessary training data for semantic annotation due to the fact that the relations and concepts used are clearly defined as well as standardized within the UNL Editor framework. The UNL Editor exhibits enormous flexibility and opportunities in handling natural language text due to the fact that it follows a linguistic framework, minding the complexity and richness of natural language.

Given a text, the editor firstly split it into sentences and thus the UNLization process follows 3 stages: (i) identification of concepts or creation the nodes (Stage 1), (ii) assigning attributes (Stage 2), and (iii) identification of relation labels between concepts (Stage 3).

In such process, we see that lexical semantics is expressed through creating the nodes, a process in which every single or compound word or rather every concept in the sentence to be analyzed is matched with its corresponding ID. In the UNLization of the sentence "*Seven people have been rescued from the rubble*" (from the English document of the cluster 09), showed in Figure 1, we identified 4 concepts in the Stage 1, codified by the following *UWs*: "7", "person", "rescue", and "rubble". Each *UW* is thus codified as a particular node in the graph. The dictionary from which the *UWs* (and IDs) are extracted is based on Princeton WordNet (version 3.0), which stores 155,287 words and expressions organized in 117,659 synsets (FELLBAUM, 1998). In order to make the process of selecting the appropriate UW easier and for more clarification to the concept, the UNL Editor provides to the annotators all information attach to each concept in WordNet, including gloss (i.e., textual description of a synset's meaning or concept) and synsets.

Grammatical or sentential semantics is expressed in Stages 2 and 3, and it is based on the assumption that the syntactic structure of the sentences overlaps with its semantics. In the UNL Editor, grammatical semantics is codified in terms of attributes and semantic relations. Codifying grammatical categories such as tense, mood, aspect, number, etc., the attributes correspond to one place predicates. They are mainly used to convey three different kinds of information: (i) role of the node in the UNL graph ('@entry', for example, indicates the main (starting) node of a UNL directed graph), (ii) grammatical knowledge conveyed by closed classes, such as affixes, determiners, adpositions, conjunctions, auxiliary and quasi-auxiliary verbs and degree adverbs, and (iii) subjectivity of sentences, i.e., what is said from the speaker's point of view, including phenomena technically called "speech acts", "propositional attitudes", "truth values", etc. In the annotation of the sentence in Figure 1, the *UW* "person", for example, received the attribute label "@pl" in Stage 2, which means that there is more than one person (plural). The *UW* "rescue" has the ALs "@past", which indicates that the event took place in the past, and "@entry", which means that this is the main UW of the sentence. The *UW* "rubble" received the attribute "@ def", which expresses definiteness and implies that "rubble" had already been mentioned before (which is expressed by the definite article "the").

For linking the concepts, the UNL Editor provides a super set of semantic relations, including 45 highly standardized labels. They are used to describe the objectivity information of the sentences. In the UNL formalism, relations are normally regarded as representations of semantic cases or thematic roles (such as agent, object, instrument, etc.) between concepts. They are used in form of arcs connecting a node to another node in a UNL graphical representation. In opposition to attributes, relations correspond to two-place semantic predicates holding between two concepts or UWs. Since there are similarities between the semantic relations and syntactic relations in name and function, it may seem that the labels used for relations are different names for special grammatical functions (ALANSARY *et al*., 2011). However, the intention is that the labels denote specific ideas rather than grammatical structures. According to Alansary *et al*. (2011), the UNL conceptual relations are more abstract than the grammatical (or syntactic) relations. In general, relations are always used to describe semantic dependencies between syntactic constituents.

For example, in the sentence "*Seven people have been rescued from the rubble*" of the Figure 1, we identified the following *RLs* in Stage 3: "qua", "obj", and "src". The binary *RL* "obj" codifies "a thing in focus which is directly affected by an event or state". In the example, "obj" links the concepts "rescue" and "person". The *RL* "qua" represents a quantity of a thing or unit. In Figure 1, "qua" interconnects the UWs "7" and "person". And, finally, "scr", which codifies "initial state, place, origin or source", is responsible for linking "rescue" and "rubble" (CHAUD, 2014).

FIGURE 1 – Sentence UNL encoding (CHAUD, 2014)

Identification of concepts/nodes (Stage 1)	**Assigning attributes (Stage 2)**	**Identification of relation (Stage 3)**
7	7	*qua*(person.@pl,7)
person	person.@pl	
rescue	rescue.@past.@entry	*obj*(rescue.@past.@entry,person.@pl)
rubble	rubble.@def	*src*(rescue.@past.@entry,rubble.@def)

3.2 The Alignment of Source Texts and Human Summaries

Many authors have used manual alignment of texts and reference summaries in Automatic Summarization, since it may reveal some of the human strategies used to produce the summary (e.g. MARCU, 1999; HIRAO; SUZUKI; ISOZAKI; MAEDA, 2004). In this particular work, the goal of the alignment was to compare sentences that were aligned to the summary to sentences that were not aligned with regard to their conceptual characteristics. As for the annotation, 1 computational linguist performed the alignment in one-hour daily sessions, during 1 month. The expert followed the methodology described in Camargo (2013). Thus, the manual alignment was performed in the summary-to-document direction and at the sentence level. Moreover, we have followed four general rules. The rule 1 specifies that a summary sentence must be aligned to a document sentence based on the content overlapping, not only considering the word overlapping between them. The rule 2 states that the alignment should first be based on the main information overlapping, i.e., the alignment should be established if the sentences express similar main topics. If this was not possible, the rule 3 establishes that a summary sentence and a document sentence may also be aligned based on secondary information overlapping. Finally, the rule 4 determines that one summary sentence should be connected to all similar (partial or total) sentences from the distinct source documents of the same cluster. Consequently, according to the rule 4, the summary-documents alignments codify one-to-many relationships. Once a summary sentence SS was linked to one or more document sentences DS, a manual correspondence between their UNL representations was also created. Figure 2 illustrates a 1:2 alignment.

In such example, the SS was aligned to two DSs because they have the same meaning or express the same topic.

FIGURE 2 – Alignment of summary and document sentences/UNL encodings

Summary sentence / UNL codification	**Source sentence / UNL codification**
Cerca de 100 pacientes tiveram que ser retirados do centro médico. [C9_ Sum-ref_S2]	Nearly 100 patients at the St John Regional Medical Center in Joplin were evacuated after the hospital took a direct hit. [C9_EN_S30]
	Pacientes tiveram que ser retirados do centro médico. [C9_PT_S9]

obj(remove.@past.@obligation.@entry,patient.@pl)
mod(center.@def,medical)
src(remove.@past.@obligation.@entry,center.@def)
qua(patient.@pl,approximately)
bas(approximately,100)
	bas(nearly,100)
	qua(patient.@pl,nearly)
	plc(patient.@pl,St John Regional Medical Center.@def)
	plc(St John Regional Medical Center.@def,Joplin)
	obj(evacuate.@past.@entry,patient.@pl)
	tim(evacuate.@past.@entry,after)
	obj(after,:01)
	aoj:01(direct,hit.@indef)
	obj:01(take.@past.@entry,hospital.@def)
	agt:01(take.@past.@entry,hit.@indef)

	obj(remove.@past.@obligation.@entry,patient.@pl)
	mod(center.@def,medical)
	src(remove.@past.@obligation.@entry,center.@def)

Table 2 shows the distribution of the different alignment types (1-n) and Table 3 describes the number of alignments where a summary sentence was aligned to source sentences(s) in just one language (Portuguese or English) or in both languages. According to the results, we may see that 8 summary sentences were aligned to only one sentence of the source texts (1-1), 7 summary sentences were aligned to 2 sentences of the source texts (1-2), and so on. The alignment illustrated in Figure 2, for example, is 1-2. From the 23 summary sentences, 15 were aligned

(65,3%) to some source sentence, with the distribution per language as described in Table 3. This result was expected, since a multi-document summary could be potentially connected to 2 related source texts of its cluster. From the 144 sentences in the source texts, 50 (37,4%) were aligned to some summary sentence, but it does not mean that the sentences were aligned only once. A sentence of a summary may be aligned to more than one sentence of the source text, and the sentences of the source texts may be redundant or even identical. Since the alignments may indicate total or partial content overlap, whenever a sentence of a given source text is aligned to a summary sentence, this means that at least part of the information conveyed by that sentence is also in the summary, indicating that the sentence brings some content considered relevant by the human summarizer. However, it is reasonable to assume that, in general, document sentences that are aligned to summary sentences carry more relevant information than sentences that are not aligned.

TABLE 2 – Alignment types in the corpus

Types of alignment	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	1:9	1:10
No. of alignments	8	7	4	0	3	0	0	0	0	1

TABLE 3 – Distribution of the alignments per language

Alignment	Summary: Portuguese	Summary: English	Summary: Both
Quantity	6	6	11

Next, we describe the conceptual measures for content selection in MMDS.

4 Lexical-Conceptual Measures

Based on the review of the literature, we have selected 3 lexical-conceptual measures that are potentially adequate to capture human content selection strategies in MMDS: (i) *concept frequency*, (ii) *concept frequency corrected by the inverted document frequency*, and (iii) *concept frequency normalized by the number of concepts in the sentence.* Given the fact our that corpus is annotated with UNL, we renamed the measures as follows: (i) simple UW frequency or F(UW); (ii) UW frequency corrected by inverse document frequency or F(UW)*IDF(UW), and (iii) UW frequency normalized to the number of UWs in the sentence or F(UW)/No. UWs in S. Considering the step (ii) of the language-dependent MMDS methods, which consists in ranking the original sentences and picking the top scoring sentences to build the multi-document extract, these three measures capture the content of a multilingual cluster by counting the occurrences of concept underlying synonyms (i.e., different words that express the same concept) and equivalences (i.e., expressions of a concept in different languages).

The selection of the F(UW) measure relies on the assumption that the most frequent concepts of a cluster express the most relevant information and, therefore, the sentences that are composed of such concepts should compose the summary. This measure has already been applied by Tosta (2014) for multilingual multi-document summarization involving the Brazilian Portuguese language (and English), only taking into account the nominal concepts of the cluster. The author showed, indeed, that selecting sentences based on conceptual knowledge rather than superficial features improves the linguistic quality of the extracts. Here, we have considered the frequency of all concepts in the input. The F(UW) equation is described in (1).

(1)

$$S(s) = \sum_{\forall UWi \in s} F(UWi)$$

where

S is the sentence scoring function;

s is the sentence being scored;

F is the concept frequency; and

UWi is the concept.

The F(UW)*IDF(UW) measure is used to evaluate how important a concept is to a document in a corpus. The importance increases proportionally to the number of times the concept appears in the document, but it is offset by its frequency in the corpus (in this case, in the 3 clusters). Thus, a higher F(UW)*IDF(UW) score indicates that a concept is important because it is frequent in the document, but relatively uncommon in the other documents of the corpus. Although F(UW)*IDF(UW) was already applied by Sornlertlamvanich et al. (2001) in automatic summarization, there are no details about the performance of this measure. Thus, we decided to explore its potential to capture human content selection preferences in MMDS. The F(UW)*IDF(UW) equation is defined in (2).

(2)

$$S(s) = \sum_{\forall UWi \in s} W(UWi)$$

$$W(UWi) = F(UWi) \quad * \quad IDF(UWi)$$

$$IDF(UWi) = \log\left(\frac{D(UWi)}{d(UWi)}\right)$$

where

s is the sentence being scored;

W is the function that calculates the score of each concept;

UWi is the concept;

F is the concept frequency;

IDF is the inverted document frequency;

D(UWi) is the number of documents of the corpus; and

d(UWi) is the number of documents in which the UW occurs.

The F(UW)/No. UWs in S measure was proposed because, according to Tosta (2014), F(UW) tends to assign better rankings to longer sentences and worse rankings to short sentences. Thus, we suggested F(UW)/No. UWs in S, which also involves calculating sentence scores based on concept frequency, but includes a normalization procedure to make sentence selection less dependent on their size. The F(UW)/No. UWs in S equation is described in (3).

(3)

$$S(s) = \frac{\sum_{\forall UWi \in s} F(UWi)}{n(s)}$$

where

S is the sentence scoring function;

s is the sentence being scored;

F is the concept frequency;

UWi is the concept; and

n(s) is the number of UWs in sentence *s*.

The application of the measures followed 3 steps: (i) calculation of the measure of each UW in the cluster, (ii) scoring all the source sentences according to the value of the measure obtained for their constitutive UWs, and (iii) ranking the sentences by their score. Thus, we built three different ranks – one for each measure.

5 Investigation of the measures for sentence selection in MMDS

Given the three different ranks, we sought to identify which of them was closer to what human summarizers did during summarization. In order to evaluate the potential of the conceptual measures (and the superficial strategy) for capturing human content selection preference, we calculated how many aligned source-sentences were covered by the top sentences of each rank. Thus, by analyzing whether these measures are capable of providing ranks in which the sentences aligned to the summary are ranked first, it is possible to evaluate whether the content selected by each measure correlates to the content selection performed by the human summarizer. Ideally, the sentences ranked first by these measures should be sentences that were aligned to the summary, because this means that they bring information related to the summary (presenting total or partial content overlap). As for low-ranking sentences, they should be non-aligned sentences, that is, they should be sentences with no relation to the summary.

In order to know how many of the top-ranked sentences were relevant based on the alignment of the human summary and source texts, we have posed the following question for each source-text: “Out of the *n* top-ranked sentences, how many were aligned to the summary?” Since the source texts vary in terms of size or length, the number of sentences (*n*) used for comparison was proportional to the text size. The *n* value was empirically defined as 20% of the number of sentences in the text, rounded down if necessary. For example, the text C1-EN has 36 sentences (Table 1), thus the 7 top-ranked sentences were used for comparison. This means that, given the 7 top-ranked sentences, we were interested in knowing how many of them had been aligned to summary sentences. We also considered a rank that was built according to the superficial *sentence* position strategy.

Table 4 shows the results of the analysis. Figure 3 shows a graphical overview of the comparison. It can be seen that the measures have similar performances.

FIGURE 3 – Graphical comparison of the relevance strategies

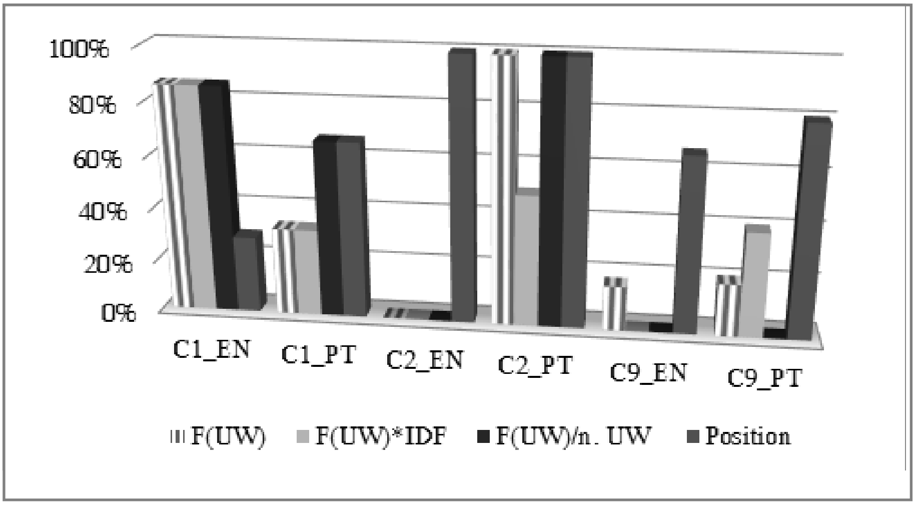

TABLE 4 – Comparison of concept-based and superficial relevance strategies

Source text	**F(UW)**		**F(UW)*IDF(UW)**		**F(UW)/No. UWs in S**		**Position**	
	Qt.	%	Qt.	%	Qt.	%	Qt.	%
C1_EN	6/7	86%	6/7	86%	6/7	86%	2/7	29%
C1_PT	1/3	33%	1/3	33%	2/3	67%	2/3	67%
C2_EN	0/2	0%	0/2	0%	0/2	0%	2/2	100%
C2_PT	1/2	100%	1/2	50%	1/2	100%	2/2	100%
C9_EN	1/6	17%	0/6	0%	0/6	0%	2/3	67%
C9_PT	1/5	20%	2/5	40%	0/5	0%	4/5	80%

In average, we verified that the 4 methods selected 48% of aligned sentences, i.e., 48% of the sentences among the top ranked ones. Therefore, we may consider a content selection strategy as successful when more than the average of the sentences selected were aligned to the summary (i.e., presented relevant content). In this case, if we approximate the value to 50%, the concept-based method with the best performance was F(UW)/No. UWs in S, as can be seen in the Table 5.

TABLE 5 – Ranks with at least 50% of aligned sentences in the top positions

Source text	**F(UW)**	**F(UW)*IDF**	**F(UW)/No. UWs in S**	**Position**
C1_EN	Yes	Yes	Yes	No
C1_PT	No	No	Yes	Yes
C2_EN	No	No	No	Yes
C2_PT	Yes	No	Yes	Yes
C9_EN	No	No	No	Yes
C9_PT	No	No	No	Yes
TOTAL	**2**	**1**	**3**	**5**

The method F(UW)*IDF was the one that led to the lowest number of aligned sentences among the top-ranked sentences. This means that, in our case, it would select very few sentences carrying content that was considered relevant by the human summarizer. In only 1 of the 6 source-texts the percentage of aligned sentences among the top-ranked ones was higher than 50% for this method. It is very hard to pinpoint the specific reasons for this result. However, the size of our corpus and the very rationale of the formula for sentence ranking seem to be relevant factors. According to the F(UW)*IDF equation (2), concepts that occur in all texts end up with weight equal to zero, which would be a way of decreasing the influence of the most common words in the language. However, in a small corpus, with 2 or 3 texts, for example, the chance that a UW occurs in every text is still relatively high, and this way of calculating the importance of a UW would assign weight zero for such UWs, therefore often disregarding important concepts.

The performance of F(UW) and F(UW)/No. UWs in S was slightly higher than that of F(UW)*IDF, although it is difficult to establish the actual significance of this difference, given our small corpus. In 3 (out of the 6) source-texts, the F(UW)/No. UWs in S measure was capable of generating ranks with more than 50% of aligned sentences among the top-ranked sentences. This means that, in half of the texts, there was good correlation between the content considered relevant by the human summarizer and the content of the sentences selected by the measure. The F(UW) measure produced ranks with at least 50% of aligned sentences in the top positions in 2 texts.

If we take a more pessimistic/rigid view and consider that a method should select 80% of the aligned sentences, the measures F(UW) and F(UW)/No. UWs in S perform equally (see Table 4).

Comparing the three concept-based relevance measures to the superficial strategy, we can see that, in 5 of the 6 texts, selecting content based on sentence position led to ranks in which the top-ranked sentences were aligned in more than 50% of the cases. In other words, in 5 out of the 6 texts, more than half of the sentences selected based on sentence position brought relevant content. It is not totally surprising that the sentence position strategy, particularly with a journalistic corpus, better captures the human preferences. Camargo *et al*. (2015) showed that, in a (monolingual) multi-document scenario, *position* is one of the main features that characterize the sentences usually selected by humans to compose a news summary. Our results seem to indicate that the first sentences of the texts did not necessarily contain the most frequent concepts of the cluster. In several cases, the sentences with the most frequent concepts were in the middle or at the end of the text.

6 Final Remarks

To the best of our knowledge, this integrated study of statistical relevance measures over a multilingual multi-document corpus annotated with UNL is new in the field of NLP, at least for the processing of Portuguese.

With regard to the potential of the conceptual-based measures, we highlight that the best performance of the superficial strategy is something worth noting. This is an interesting result because it may reflect dissociation between sentences located in the beginning of the text and sentences with the most frequent concepts. Throughout the corpus, very often it was noticed that sentences in intermediate or final position in the text were the ones bringing the most frequent concepts of the cluster. If the fact that a text belongs to the journalistic genre means that its first sentences bring the most relevant information, and if its first sentences do not necessarily contain the most frequent concepts (as suggested in this study), one can conclude that a relevant sentence is not necessarily a sentence bringing the most frequent concepts of the cluster. Therefore, the assumption that relevant concepts tend to appear repeatedly throughout the cluster perhaps has to be reassessed, or at least applied with some caution. It is important to keep in mind that this was a small-scale study and, therefore, definitive conclusions or generalizations should be avoided.

Future work may include the study of the measures using a bigger news corpus or a data collection of a different genre, especially one in which sentence position would not be a feature so important to indicate content “relevance”. Of course, these extensions will require semantic annotation of the corpora, which is a complex and time-consuming (semiautomatic) task, but necessary for future advances in the field. Another possibility is to use more than one manual (or reference) summary to evaluate the potential of the metrics, since summarization is a very subjective task and different reference summaries could reveal different content strategies. Moreover, future work may include the production of automatic summaries based on the ranks and the manual evaluation of their linguistic quality following criteria such as those that were used in DUC.

In addition to allowing deeper investigation on concept-based measures, a larger corpus annotated with UNL could provide the data necessary to explore abstractive MMDS strategies, such as those proposed by Sornlertlamvanich *et al*. (2001) (e.g., combining sentences that cover the same concepts) for single-document summarization.

Acknowledgements

The authors thank Coordination for the Improvement of Higher Education Personnel CAPES, CNPq, and State of São Paulo Research Foundation (FAPESP) for the financial support.

References

ALANSARY, S.; NAGI, M.; ADLY, N. UNL Editor: An annotation tool for semantic analysis. In: INTERNATIONAL CONFERENCE ON LANGUAGE ENGINEERING, 11., 2011, Cairo, Egypt. *Proceedings...* Cairo, Egypt, 2011.

CAMARGO, Renata. T. *Investigação de estratégias de sumarização humana multidocumento*. 2013. 133 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2013.

CAMARGO, R.T.; DI FELIPPO, A.; PARDO, T.A.S. On strategies of human Multi-Document Summarization. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 10, 2015, Natal, Brazil. *Proceedings...* Natal, 2015, p. 141-150.

CARDEÑOSA, J. *et al*. A new knowledge representation model to support multilingual ontologies. A case study. In: INTERNATIONAL CONFERENCE ON SEMANTIC WEB AND WEB SERVICES (SWWS), 2008, Monterrey, Mexico. *Proceedings...* Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 313-319. ISBN 1-60132-089-2.

CHAUD, M. R. Investigação de estratégias de Sumarização Automática Multidocumento Multilíngue baseadas em interlíngua. 2014. 100f. Qualificação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2014.

DANG, H. T. Overview of DUC 2005. In: DOCUMENT UNDERSTANDING CONFERENCE, 2005.

DI-FELIPPO, A. CM2News: Towards a Corpus for Multilingual Multi-document Summarization. In: CORPORA AND TOOLS FOR PROCESSING CORPORA WORKSHOP (CTPC) (PROPOR), 12., 2016, Tomar, Portugal. *Proceedings...* Tomar, July, 2016.

DI-FELIPPO, A.; TOSTA, F. E. S.; PARDO, T.A.S. Applying Lexical-Conceptual Knowledge for Multilingual Multi-Document Summarization. In: INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR), 12., 2016, Tomar, Portugal. *Proceedings...* Tomar, July, 2016. Doi: https://doi.org/10.1007/978-3-319-41552-9_4.

EVANS, D. K., KLAVANS, J. L. Columbia Newsblaster: multilingual news summarization in the web. In: HUMAN LANGUAGE TECHNOLOGIES (HLT) – NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (NAAL), 2004. Boston, MA: NAACL, 2004. Doi:10.3115/1614025.1614026.

EVANS, D. K., MCKEOWN K.; KLAVANS, J. L. Similarity-Based Multilingual Multi-document Summarization. *Technical Report CUCS-014-05*, Columbia University, New York, 2005. ISBN-13: 978-0262061971 ISBN-10: 026206197X.

FELLBAUM, Christiane D. (Ed.). *Wordnet:* an electronic lexical database. Massachusetts: MIT Press, 1998.

HENNIG, L., UMBRATH, W., WETZKER, R. An ontology-based approach to text summarization. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING AND ONTOLOGY ENGINEERING (NLPOE 2008), 3., Toronto, 2008. *Proceedings...* Toronto, Canada, 2008. p. 291-294. Doi: https://doi.org/10.1109/WIIAT.2008.175.

HIRAO, T.; SUZUKI, J.; ISOZAKI, H.; MAEDA, E. Dependency-based Sentence Alignment for Multiple Document Summarization. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING), 2004. *Proceedings...* Switzerland, 2004. p. 446-452. Doi: https://doi.org/10.3115/1220355.1220419.

KIM, S. N.; BALDWIN, T.; KAN, M.-Y. Extracting domain specific words – a statistical approach. In: AUSTRALASIAN LANGUAGE TECHNOLOGY ASSOCIATION WORKSHOP, 2009. *Proceedings...* Sidney, Australia, 2009. p. 94-98.

KIT, C.; LIU, X. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*. International Journal of Theoretical and Applied Issues in Specialized Communication, John Benjamin Publishing Company, v.14, n. 2, p. 204-229, 2008. Doi: 10.1075/term.14.2.05kit.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. *Applied Soft Computing*, v. 12, n. 10, p. 3124-3131, 2012. Doi: https://doi.org/10.1016/j.asoc.2012.06.017.

LIN, C-Y. ROUGE: a Package for Automatic Evaluation of Summaries. In: WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT (WAS), 2004, Barcelona. *Proceedings*... Barcelona, 2004.

LOPES, L.;FERNANDES, P.; VIEIRA, R. Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. *Knowledge-Based Systems*, Elsevier, v. 97, p. 237-249, 2016. Doi: https://doi.org/10.1016/j.knosys.2015.12.015.

MANGAIRKARASI, Selvi; GUNASUNDARI, Salem. Semantic based text summarization using universal networking language. *International Journal of Applied Information System*, New York, v.3, n.8, p. 18-23, 2012. ISSN: 2249-0868.

MARCU, Daniel. The automatic construction of large-scale corpora for summarization research. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999. *Proceedings*... 1999, p. 137-144. Doi: 10.1145/312624.312668.

MARTINS, Camila. B. *UNLSumm:* Um sumarizador automático de textos UNL. 2002. 100 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, SP, 2002.

MARTINS, Camila B.; RINO, Lucia H. M. Heurísticas de poda de sentenças para a Sumarização Automática de textos UNL: Estudo de casos. *Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC – ICMC* (Relatório NILC-TR-02-11). São Carlos, SP: USP, 2002. 51p.

MARTINS, R. T. *et al*. *The UNL distinctive features: evidences through a NL-UNL encoding task*. In: INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION CONFERENCE (The First International Workshop on UNL, Other Interlinguas and Their Applications), 2002. *Proceedings*... Las Palmas, 2002. p. 8-13.

MCKEOWN, K. R. *et al.* Tracking and summarizing news on a daily basis with columbia's newsblaster. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH (HLT´02), 2, 2002, San Diego, USA. *Proceedings...*San Diego, 2002, p. 280-285. Doi: https://doi.org/10.3115/1289189.1289212.

MORATO, J. *et al*. Wordnet applications. In: INTERNATIONAL GLOBAL WORDNET CONFERENCE, 2., 2004. *Proceedings...* Masaryk University, Brno, 2004. p. 270-278. ISBN 80-210-3302-9.

ORĂSAN, C.; CHIOREAN, O. A. Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC), 6., 2008, *Proceedings...* Marrakesh, 2008. p. 2114-19.

PANDIAN. L. S.; KALPANA. S. UNL based Document Summarization based on Level of Users. *International Journal of Computer Applications,* New Yor, v. 66, n. 24, p. 28-36, March 2013.

ROARK, B.; FISHER, S. OGI / OHSU baseline multilingual multi-document summarization system. In: MULTILINGUAL SUMMARIZATION EVALUATION (MSE) (Association for Computational Linguistics Workshop), 2005, Michigan, United States of America. *Proceedings...* Michigan, USA, 2005.

SARKAR, K.; BANDYOPADHYAY. S. A multilingual text summarization system for Indian languages. In: SYMPOSIUM ON INDIAN MORPHOLOGY, PHONOLOGY & LANGUAGE ENGINEERING (SIMPLE'05), 2., 2005, Kharagpur, India. *Proceedings...* Kharagpur: Indian Institute of Technology, 2005. February 5-7.

SARKAR, K. Multilingual summarization approaches. In: *Computational Linguistics*: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications. Information Resources Management Association, 2014. p. 158-177. Doi: https://doi.org/10.4018/978-1-4666-6042-7.ch009.

SORNLERTLAMVANICH, V.; POTIPITI, T.; CHAROENPORN, T. UNL document summarization. In: INTERNATIONAL WORKSHOP ON MULTIMEDIA ANNOTATION (MMA'2001), 1., 2001, Japan. *Proceedings...* Tokyo, Japan, 2001.

TOSTA, F. E. S.; DI-FELIPPO, A.; PARDO, T. A. S. Estudo de métodos clássicos de sumarização no cenário multidocumento multilíngue. In: STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (TILiC), 3., 2013. Fortaleza. *Proceedings...* Fortaleza: Sociedade Brasileira de Computação, 2013.p.1-3.

TOSTA, F. E. S. Aplicação de conhecimento léxico-conceitual na sumarização automática multidocumento multilíngue. 2014. 116 f. Dissertação (Mestrado em Linguística), Universidade Federal de São Carlos, São Carlos, SP, 2014.

UCHIDA, H.; ZHU, M.; DELLA SENTA, T. *The UNL, a Gift for a Millennium*. Tokyo, Japan: The United Nations University - Institute of Advanced Studies, 1999.

WAN, X.; LI, H.; XIAO, J. Cross-language document summarization based on machine translation quality prediction. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala, Sweden. *Proceedings*... Stroudsburg, PA: Association for Computational Linguistics, 2010. p. 917-926.

WU, Chia-Wei; LIU, Chao-Lin. Ontology-based text summarization for business news articles. In: INTERNATIONAL CONFERENCE ON COMPUTERS AND THEIR APPLICATIONS (ISCA), 2003, Hawaii, USA. *Proceedings...* Hawaii, 2003. p. 389-392.