CM2News: Towards a Corpus for Multilingual Multi-Document Summarization

Ariani Di-Felippo[1,2]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos/SP, Brazil
Av. Trabalhador São-carlense, 400, São Carlos, 13566-590, Brazil
[2] Language and Literature Department, Federal University of São Carlos (UFSCar)
Rodovia Washington Luís, km 235 - SP 310, São Carlos, 13565-905, Brazil
arianidf@gmail.com

Abstract. This paper describes the ongoing construction of CM2News, a semantic-annotated corpus for fostering research on multilingual multi-document summarization. The corpus comprises 20 clusters of news texts in English and Brazilian Portuguese languages and a set of multi-document manual and automatic summaries. All the source texts have a layer of semantic annotation at lexical level. Some clusters also have annotation at sentence level, as well as alignment of texts and human summaries. The corpus is a result delivered within the context of the *Sustento* Project, which aims at generating linguistic knowledge for multi-document summarization. The corpus design and the manual annotation tasks are detailed in this paper.

Keywords: corpus, multilingual resource, multi-document summarization.

1 Introduction

As the amount of on-line news texts in different languages is growing at an exponential pace, Multilingual Multi-Document Summarization (MMDS) is a quite desirable task. It aims at identifying the main information in a cluster of (at least) two news texts, one in the user's language and one in a foreign language, and presenting it as a coherent/cohesive summary in the user's languages.

The ongoing Sustento project[1] tackles this and also other multi-document summarization tasks. Specifically, it has been focusing on 3 correlated tasks: (i) characterization of the human multi-document summarization (HMS) and development of automatic methods based on HMS strategies, (ii) study of the multi-document phenomena (e.g., redundancy) and proposition of methods for their automatic detection, and (iii) development of deep methods based on semantic-conceptual representation of the source texts. The project is mainly corpus-driven, i.e., linguistic descriptions, tools and applications are drawn upon corpora. This motivates our interest in constructing CM2News, a *Multi-document Bilingual Corpus of News Texts* for MMDS, which was first described in [1].

[1] http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23

The CM2News comprises 20 clusters of news texts. Each cluster is composed of 2 source-texts, 1 in English (En) and 1 Brazilian Portuguese (BP), and a set of multi-document manual and automatic summaries. Given our interest in exploring deep summarization based on semantic-conceptual knowledge, all the source texts were manually annotated using Princeton WordNet [2], and some clusters were also annotated following UNL (*Universal Network Language*) formalism [3]. We also carried out the sentential alignment of texts and human summaries of some clusters based on overlapping content between the sentences.

To the best of our knowledge, CM2News is the first multi-document corpus with multilingual clusters that include Portuguese. This paper focuses on its manual annotation in order to produce a resource for MMDS. Section 2 first reports the corpus design. Section 3 focuses on the annotation tasks, which include the meaning representation following two different conceptual models, and the alignment of texts and human summaries. In Section 4, we briefly highlight the projects that already made use of CM2News. Section 5 provides some final remarks and future works.

2 Building Principles

According to [4], a well-designed corpus should reflect its purpose. Since our corpus has been building for MMDS, it is a multi-document and multilingual resource. This means that its internal structure is based on *clusters*, and each cluster is composed of texts in different languages on the same topic. The CM2News corpus has 20 clusters, and each of them is composed of 2 news texts, 1 in En and 1 in BP. The corpus sums up 40 texts altogether, amounting to 19.984 words.

The texts in En and BP were manually collected from the *BBC*[2] and *Folha de São Paulo*[3] on-line news agencies, respectively. To collect them, we have followed the 3 criteria that were applied to build CSTNews [5], a reference corpus in BP for Multi-Document Summarization (MDS). One criterion was to collect texts with similar length (in terms of words). For example, the texts in En and BP of the cluster C19 have 446 and 452 words, respectively. Another criterion was selecting topics with high popularity on the web, which means that CM2News only cover trending topics at the time of the corpus construction (e.g., "Angelina Jolie's mastectomy" in 2013). Finally, according to the diversity guideline, the clusters cover a variety of domains, i.e., world (8 clusters), politics (3 clusters), health (4 clusters), science (3 clusters), entertainment (1 clusters), and environment (1 clusters). Moreover, each cluster of our corpus also has 1 human multi-document *abstract*[4], and automatic multi-document extracts generated by baseline and deep MMDS methods. Both human and automatic summaries are written in Portuguese, but they are ideally brief representations of the essential content of the two source-texts. All summaries were generated based on a compression rate of 70%, which means that they correspond to 30% of the size of the largest text of the cluster.

The next section describes the linguistic annotations of the CM2News corpus.

[2] http://www.bbc.co.uk/news/

[3] http://www.folha.uol.com.br/

[4] Abstracts are summaries that contain some degree of paraphrase of the input.

3 Corpus Annotation

3.1 Lexical Semantic Annotation

The 40 source texts of the CM2News corpus have a layer of semantic annotation at lexical level. Specifically, the common nouns, which cover part of the main content of a text, were semi-automatically tagged with their correspondent concept.

In order to identify the nominal concepts in the texts, we made use of WordNet[5] lexical database. Although WordNet's fine-grained senses may create difficulties for annotating nouns, we have chosen such database due to its widespread application in several NLP tasks and broad coverage, and the still partial development of similar resources for Portuguese.

The annotation was carried out by groups of 2 or 3 experts[6], in a total of 12 computational linguists, in daily meetings from 90 to 120 minutes. The annotation process, including 1 day for training, took the period of 15 days. For each cluster to annotate, the experts were organized in different groups, trying to avoid any annotation bias. To assist the experts, we built an easy-to-use editor called MulSen[7] (Multilingual Sense Estimator). Given a cluster, the editor first performs an automatic pre-processing step, which consists in annotating the source texts with part-of-speech (POS) tags. MulSen incorporates two taggers, one for each language, and the output of such tools can be manually revised if necessary. Once the texts are tagged, the annotation of a noun *n* in Portuguese, in particular, starts with the automatic translation of *n* to English, since WordNet codifies the concepts by sets of synonyms in English. The translation is performed using the online bilingual dictionary WordReference®[8], but the editor also allows the manual inclusion of a translation equivalence. Finally, the editor suggests the best synset that represents the underlying concept of *n,* which should be validated by the experts to complete the process. The suggestion results from the application of a word sense desambiguation algorithm [6]. If the suggested synset is not appropriate, the editor displays all the synsets containing the English translation of *n* and then the annotators are able to select a more suitable option among them. The annotation of a text in English basically follows the same procedure except the machine-translation stage.

The experts have followed 4 general rules in order to annotate the nouns: (i) firstly annotate the text in English of a cluster, since its vocabulary can provide appropriate translations for the annotation of the nouns in Portuguese, (ii) annotate the POS silence, i.e., nouns that were not automatically detected, (iii) ignore the POS noise, i.e., words that were wrongly annotated as nouns, and (iv) annotate every occurrence of a concept (i.e, synonyms and equivalences) in the cluster with the same (and more adequate) synset.

---

[5] A semantic network of English in which the meanings of words and expressions of noun, verb, adjective, and adverb classes are organized into "sets of synonyms" (*synsets*). Each *synset* expresses a distinct concept and they are interlinked through conceptual-semantic (hyponymy, meronymy, entailment, and cause) and lexical (antonymy) relations [2].

[6] The agreement rate has not been calculated yet.

[7] http://www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html

[8] http://www.wordreference.com/

The annotation was also performed according to 4 specific rules. Since the taggers only detect single word forms, the first rule establishes that every common noun that is a multiword expression head should be annotated with a synset that codifies the expression's sense. For instance, the head (shown in italics) of the multiword expression "*gás* de pimenta" ("pepper spray") was annotated with the *synset* {pepper spray} ("a nonlethal aerosol spray made with the pepper derivative oleoresin capiscum"). Following this rule, we were able to encode complex concepts by annotating single words only. The second rule determines that the annotators should analyze all the possible translations provided by WordReference before selecting one. This is particularly important because the adequate translation may not be the first option in the list of equivalences provided by the editor. The same procedure should be followed regarding the synset selection. When the editor suggests an inadequate synset, the annotators should carefully analyze the other options retrieved from the database. For cases where translations have to be manually inserted in the editor, the third rule establishes that the annotators should look for equivalences in external resources (e.g., *Google Translator*®[9], *Linguee*[10], and other dictionaries) and analyzes all synsets retrieved from WordNet by testing the equivalences. The forth rule determines that, if a specific concept is not covered by WordNet, it should be selected a more general one. This means that, if any of the synsets retrieved by the chosen translation is adequate, the annotators should look for a satisfactory hypernym synset.

The example (1) provides an illustration of an annotated sentence. The 4 nouns (shown in bold) that occur in the English sentence "Brazil's opening Confederations Cup match was affected by protests that left 39 people injured" (C17) were tagged with the correspondent synset, indicated between "{}". For a better comprehension, we provide the gloss (i.e., an information definition of the concept) of each synset.

(1) Brazil's **opening**<{opening} "a ceremony accompanying the start of some enterprise"> Confederations Cup **match**<{match} "a formal contest in which two or more persons or teams compete"> was affected by **protesters**<{dissenter, dissident, protester, objector, contestant} "a person who dissents from some established policy"> that left 39 **people**<{people} "any group of human beings"> injured.

3.2 Sentential Semantic Annotation

Besides the semantic annotation at lexical level, some clusters were also annotated at sentential level[11], a task first described by [7]. Both source texts and human summaries were annotated with the UNL [10] formalism, in a process called UNLization. UNL is aimed at expressing information conveyed by natural language (NL) sentences through binary relations between concepts [7]. Thus, UNL is not different from the other formal languages devised to represent NL sentence meaning [8]. The general syntax of the relations is RL(UW1,UW2), where RL stands for a Relation Label, which signals the semantic relation, and UWn, for Universal Words, which signal the related concepts. RLs are specified through mnemonics, for example, *agt* for *agent*, *mod* for *modifier*, or *obj* for *object*. UWs, in particular, constitute the

[9] https://translate.google.com/
[10] http://www.linguee.com/
[11] There is no connection between the lexical and sentential annotations so far.

UNL vocabulary, and can 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 [9], the advantages of UNL are: (i) flexibility and neutrality, since it is a language created to represent any content in any domain in any language, and (ii) generality, since the set of UWs[12] and *RLs* is sufficient to describe any kind of content expressed in NLs.

From the 20 clusters, three (C1, C2, and C9) were annotated with UNL, in a total of 158 sentences (3504 words). Each cluster was manually tagged with the support of the UNL Editor [10]. One computational linguist carried out the task in two-hours daily sessions, during 3 months. Given a text, the editor first split it into sentences and then the UNLization follows 3 stages: (i) identification of concepts (Stage 1); (ii) assigning attributes (Stage 2), and (iii) identification of relations between concepts (Stage 3). The UNLization of the English sentence "*Seven people have been rescued from the rubble*" is shown in Figure 1. In Stage 1, we identified 4 UWs making use of the dictionaries available in the editor: "7", "person", "rescue", and "rubble". In Stage 2, the UW "person" received the attribute label "@pl", which means that there is more than one person (plural) involved in the event. The UW "rescue" has two 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. In Stage 3, three RLs were identified: "qua" (quantity), "obj" (affected thing), and "src" (source). The binary RL "qua", for example, interconnects the UWs "7" and "person". Next, we describe the manual alignment of source texts and human summaries.

Stage 1	Stage 2	Stage 3
7 person rescue rubble	7 person.@pl rescue.@past.@entry rubble.@def	*qua*(person.@pl,7) *obj*(rescue.@past.@entry,person.@pl) *src*(rescue.@past.@entry,rubble.@def)

Fig. 1. Sentence UNL encoding.

3.3 Alignment of Abstracts and News Texts

Many authors have been using manual alignment of texts and reference summaries in Automatic Summarization, since it may reveal some of the human strategies used to produce the summary [11], [12]. Thus, one computational linguist has performed the alignment in one-hour daily sessions, during 1 month. The expert has followed the methodology described in [13] to align 3 clusters (C1, C2 and C9). The manual alignment was performed in the summary-to-documents direction and at sentence level, and the links were established based on total or partial content overlap. In this multi-document setting, a summary sentence may be aligned to more than one document sentences. Once the raw sentences were linked, their correspondent UNL codifications were also connected. Figure 2 illustrates the alignment.

[12] Although UWs take their meanings from English word senses, each universal word expresses a very definite meaning so lexical ambiguity is kept to a minimum.

Summary sentence / UNL codification	Source sentence / UNL codification
Cerca de 100 pacientes tiveram que ser retirados do centro médico. (*About 100 patients had to be removed from the medical center*) [C9_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. (*Patients had to be withdrawn from the medical center*) [C9_BP_S9]
obj(remove.@past.@obligation.@entry,patient.@pl) mod(center.@def,medical) src(remove.@past.@obligation.@entry,center.@def) qua(patient.@pl, nearly) bas(nearly,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)

Fig. 2. Alignment of sentences and their correspondent UNL encodings.

In Figure 2, for example, the summary sentence S2 is aligned to the following two source sentences because they share the main information: S30 from the English text and S9 from the Portuguese text. Thus, their correspondent UNL representations were linked as well. Table 1 shows the distribution of the different alignment types (1-*n*). Table 2 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.

Table 1. Distribution of the alignment types in the corpus.

Alignment	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	1:9	1:10
Quantity	8	7	4	0	3	0	0	0	0	1

Table 2. Distribution of the alignments per language.

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

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 2. 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. Next, we give an idea on how the CM2News corpus has been used in MMDS.

4 CM2News in MMDS Projects

Using CM2News, [1] has developed two deep MMDS methods for generating extracts in Portuguese. The methods select sentences to compose extracts based on the frequency of occurrence of their nominal concepts in the cluster. To score and rank the sentences, they make use of the synset annotation. The CF (concept frequency) method selects the top-ranked sentences, independently of their source language. If a selected sentence is in English, it is automatically translated to Portuguese. The CFUL (concept frequency + user language) method is driven by the user's language. It exclusively selects the top-ranked sentences from the text written in Portuguese to compose the summary, also avoiding redundancy. In an intrinsic evaluation, the methods have outperformed a sentence position *baseline* (which applies a MT strategy over the source texts) in terms of informativiness and linguistic quality.

Using the UWs from the UNL annotation, [7] has explored 3 conceptual measures to capture relevant content in MMDS: (i) CF (concept frequency), (ii) CF*IDF (concept frequency corrected by the inverted document frequency), and (iii) CF/No. of Cs (concept frequency normalized by the number de concepts in the sentence). The author has compared the measures to a superficial *sentence position* method. To evaluate the potential of the measures in capturing human preferences, the author ranked the source sentences according to each strategy, and calculated how many aligned source-sentences were covered by the top sentences of each rank. The concept-based method with the best performance is (iii), but it does not outperform the sentence position method.

4 Future Works and Final Remarks

This paper described the linguistic annotation of the CM2News corpus, which aims at supporting the investigation of deep strategies on MMDS involving Portuguese. The corpus and tools are all available on the *Sustento* Project website. We hope CM2News may foster research not only on summarization and semantic analysis, but also in other Natural Language Processing areas. Future work includes increasing the quantity of clusters, extending the UNL annotation to the entire corpus, and annotating other kinds of lexical concepts, as those expressed by verbs, for example.

Acknowledgments. We thank CNPq (#483231/2012-6), and FAPESP (#2012/13246-5) for the financial support.

References

1. Tosta, F.E.S.: Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue. 2013. Dissertação (Mestrado em Linguística) - Departamento de Letras, Universidade Federal de São Carlos (2014)
2. Fellbaum, C. (Ed.): Wordnet: an electronic lexical database (Language, speech and communication). Massachusetts: MIT Press (1998)
3. Uchida, H., Zhu, M.; Della Senta, T.: The UNL, a Gift for a Millennium. The United Nations University - Institute of Advanced Studies, Tokyo, Japan (1999)
4. Sinclair, J.: Corpus and Text - Basic Principles. In: Wynne, M. (Ed.). Developing Linguistic Corpora: a Guide to Good Practice, Oxbow Books, pp. 1-16 (2005)
5. Cardoso, P.C.F., Maziero, E.G., Jorge, M.L.C., Seno, E.M.R., Di-Felippo, A., Rino, L.H.M., Nunes, M.G.V., Pardo, T.A.S.: CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: 3rd RST Brazilian Meeting, pp. 88-105. Cuiabá, MT, Brazil (2011)
6. Nóbrega, F.A.A.: Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - ICMC, USP, São Carlos (2013)
7. Chaud, Matheus. R. Investigação de estratégias de seleção de conteúdo baseadas na UNL (*Universal Networking Language*). 2015. 157 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2015.
8. Martins, R. T. et al.: The UNL distinctive features: evidences through a NL-UNL encoding task. In: 1rst International Workshop on UNL, other Interlinguas and their Applications, LREC, 2002. Las Palmas, pp. 08-13. (2002)
9. 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), pp. 313-319. Springer Berlin Heidelberg, Berlin (2008)
10. Alansary, S., Nagi, M., Adly, N.: UNL Editor: An annotation tool for semantic analysis. In: 11th International Conference on Language Engineering. Cairo, Egypt (2011)
11. Marcu, D.: The automatic construction of large-scale corpora for summarization research. In: 22th Conference on Research and Development in IR, pp.137-44 (1999)
12. Hirao, T., Suzuki, J., Isozaki, H., Maeda, E.: Dependency-based Sentence Alignment for Multiple Document Summarization. In: International Conference on Computational Linguistics (COLING). Switzerland, pp. 446-452 (2004)
13. Camargo, R.T., Di Felippo, A., Pardo, T.A.S.: On Strategies of Human Multi-Document Summarization. In: 10th Brazilian Symposium in Information and Human Language Technology - STIL, pp. 141-150. Natal, Brazil (2015)