Summarizing through sense concentration and Contextual Exploration rules: the CHORAL system at TAC 2009

Jorge García Flores Olivier Ferret
Gaël de Chalendar

Institut de Radioprotection et de Sûreté Nucléaire
DS/DICST
31, avenue de la Division Leclerc
92260 Fontenay aux Roses, France
jorge.garcia-flores@irsn.fr

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Fontenay-aux-Roses, F-92265, France.
{olivier.ferret,gael.de-chalendar}@cea.fr

Abstract

This paper presents LIC2M's second participation in TAC evaluation campaign (Update Summarization task). Two runs were submitted: simple summarization through sense concentration and combined summarization through sense concentration and Contextual Exploration rules. The sense concentration feature is based on the unsupervised recognition of word senses from a large corpus. Contextual Exploration feature is based on journalistic discourse analysis, which results on a set of rules in order to annotate meaningful sentences in press releases.

1 Introduction

The goal of the CHORAL[1] system is to summarize scientific documents related to nuclear energy. This paper presents the modifications made to CHORAL for the TAC 2009 campaign and its results. CHORAL follows an extractive approach: the summary is built from literal sentences of the source documents. Two features are used to weight document sentences: sense concentration and cue word oriented rules (we call them Contextual Exploration rules).

The use of sense concentration aims at identifying those sentences that concentrate the main document topics. Sense concentration is calculated by relying on frequent co-occurrents in a large corpus of press articles: two years of the Los Angeles Times (LA Times) newspaper for English and two years of the Le Monde newspaper for French [Fer04]. For scientific related documents, we built senses from 6,000 scientific articles of the French Atomic Energy Commission database. For TAC 2009, only the LA Times base of senses was used.

Contextual Exploration's goal [Des06] is to select relevant sentences according to shallow linguistic markers that are likely to indicate pertinence in journalistic discourse. Markers of novelty (*the newest*, *the oldest*), uniqueness (*the one and only*), abundance (*too much*), rarity or lack were analyzed in order to get a set rules. The underlying hypothesis is that a set of linguistic marks of "extraordinarity" can indicate a relevant sentence from a journalistic perspective. These rules are used as a complementary feature of sense concentration.

Semantic resources and frequency analysis have been widely used for automatic summarization [NV05, CG98]. Semantic resources go from Wordnet [HL98] to domain-specialized thesaurus [RHB07]. Contextual exploration

[1] CHaine d'Outils pour le Résumé Automatique du LVIC: pipeline of automatic summarization tools of the LVIC laboratory.

was first applied to automatic summarization in [Bla08]. CHORAL is the first approach that combines Contextual Exploration with other features.

The paper is organized as follows: section 2 describes sense bases, which are our main semantic resource for sense concentration weighting; section 3 presents our summarization features; section 4 analyzes the scores of our systems, section 5 presents an alternative evaluation of sense-based extraction and finally, section 6 presents some conclusions and further work.

2 Word senses

The word senses we use in this work were built according to the method described in [Fer04]. More precisely, the building process starts from a network of lexical co-occurrences extracted from a corpus. First, the subgraph of the co-occurrents of each target word is delimited and turned into a similarity graph where the similarity between two co-occurrents is equal to their cohesion in the network. Then, a clustering algorithm is applied for detecting high-density areas in this graph. Finally, a word sense is defined from each resulting cluster.

mouse-device	computer#n, disk#n, pc#n, software#n, user#n, machine#n, screen#n, compatible#a ...
mouse-animal	hormone#n, tumour#n, immune#a, researcher#n, animal#n, disease#n, gene#n ...

Table 1: Two discriminated senses of the word “mouse”

An example of such word senses is given by Table 1 with the two senses found for the noun *mouse*.

3 Sense concentration and Contextual Exploration rules

The two versions of CHORAL that participated to TAC 2009, ceaList25 and ceaList31, perform the same semantic analysis, where each noun, verb and adjective from the source text is submitted to the sense base to access its different meanings according to a co-occurrence analysis from a large corpus (see Section 2). The result of this analysis is a set of sentences where the most relevant senses from the source document are present.

System ceaList25 performs an extra-step by applying Contextual Exploration rules. For each run, documents of set A and set B were considered differently. Documents from set A were processed using an unlimited sense concentration approach while documents from set B were treated with a limited sense concentration approach.

The basic operation pipeline for both systems, as presented in [GFdC08], is:

1. Multiple document fusion
2. Morphosyntactic analysis and segmentation using LIMA analyzer [BdCF$^+$ar]
3. Semantic analysis
4. Sense concentration weighting
5. Contextual Exploration rules (only for ceaList25).

3.1 System ceaList31 set A: unlimited sense concentration

Given a word lemma w, $frequency(w)$ is the number of times that w appears in the source text. $\{S_{w,1}, S_{w,2}, ..., S_{w,n}\}$ represents the set of w's senses. The relevance of a sense $S_{w,i}$ is calculated by normalizing its frequency by the frequency of all the word senses found in the source text. The relevance of w is then given by the relevance of its predominant sense and the sentence relevance is the sum of the relevance of each word, normalized against the sentence size.

$$frequency(S_{w,i}) = \sum_{u \in S_{w,i}} frequency(u)$$

$$relevance(S_{w,i}) = \frac{\sum_{u \in S_{w,i}} frequency(u)}{\sum_{t \in \{source\_text\}} \sum_{x \in S_t} frequency(x)}$$

$$relevance(w) = \underset{i}{\operatorname{argmax}}\ relevance(S_{w,i})$$

$$relevance(sentence) = \frac{\sum_{w \in sentence} relevance(w)}{|\{w \in sentence\}|}$$

3.2 System ceaList31, set B: question-limited sense concentration

Given two word lemmas, w and q, the semantic intersection between q and w's respective senses is represented by

$$S_{w \cap q} = \{S_{w,1}, S_{w,2}, ..., S_{w,n}\} \cap \{S_{q,1}, S_{q,2}, ..., S_{q,n}\}$$

where q represents the words of the topic statement and $S_{w \cap q}$, the intersection of the senses issued from q and those from any word w from the source text.

$$frequency(S_{w \cap q,i}) = \sum_{u \in S_{w \cap q,i}} frequency(u)$$

$$relevance(S_{w \cap q,i}) = \frac{\sum_{u \in S_{w \cap q,i}} frequency(u)}{\sum_{t \in \{source\_text\}} \sum_{x \in S_t} frequency(x)}$$

$$relevance(w \mid q) = \underset{i}{\operatorname{argmax}}\ relevance(S_{w \cap q,i})$$

$$relevance(sentence) = \frac{\sum_{w \in sentence} \sum_{q \in question} relevance(w \mid q)}{|\{w \in sentence\}|}$$

3.3 System ceaList25: Contextual Exploration rules

System ceaList25 combines the sense concentration feature presented above (unlimited for set A, limited for set B) with Contextual Exploration rules for the semantic values considered as relevant for journalistic discourse by [Whi06], like novelty (*the newest, the oldest*), uniqueness (*the one and only*), abundance (*too much*), rarity or lack (see Figure 1).

Contextual Exploration rules for discourse analysis were presented in [Des06] and were first applied to automatic summarization by [Bla08]. Contextual Exploration assumes that, in an information retrieval process performed by a human, the reader focuses first on structural and discourse units. These units are called markers and that trigger an examination of textual context surrounding the marker in order to confirm an information retrieval hypothesis.

Figure 1 shows the manual process for building Contextual Exploration rules:

1. Linguistic analysis of a representative corpus
2. Semantic value attribution
3. Regular expression analysis to find markers
4. Rule definition.

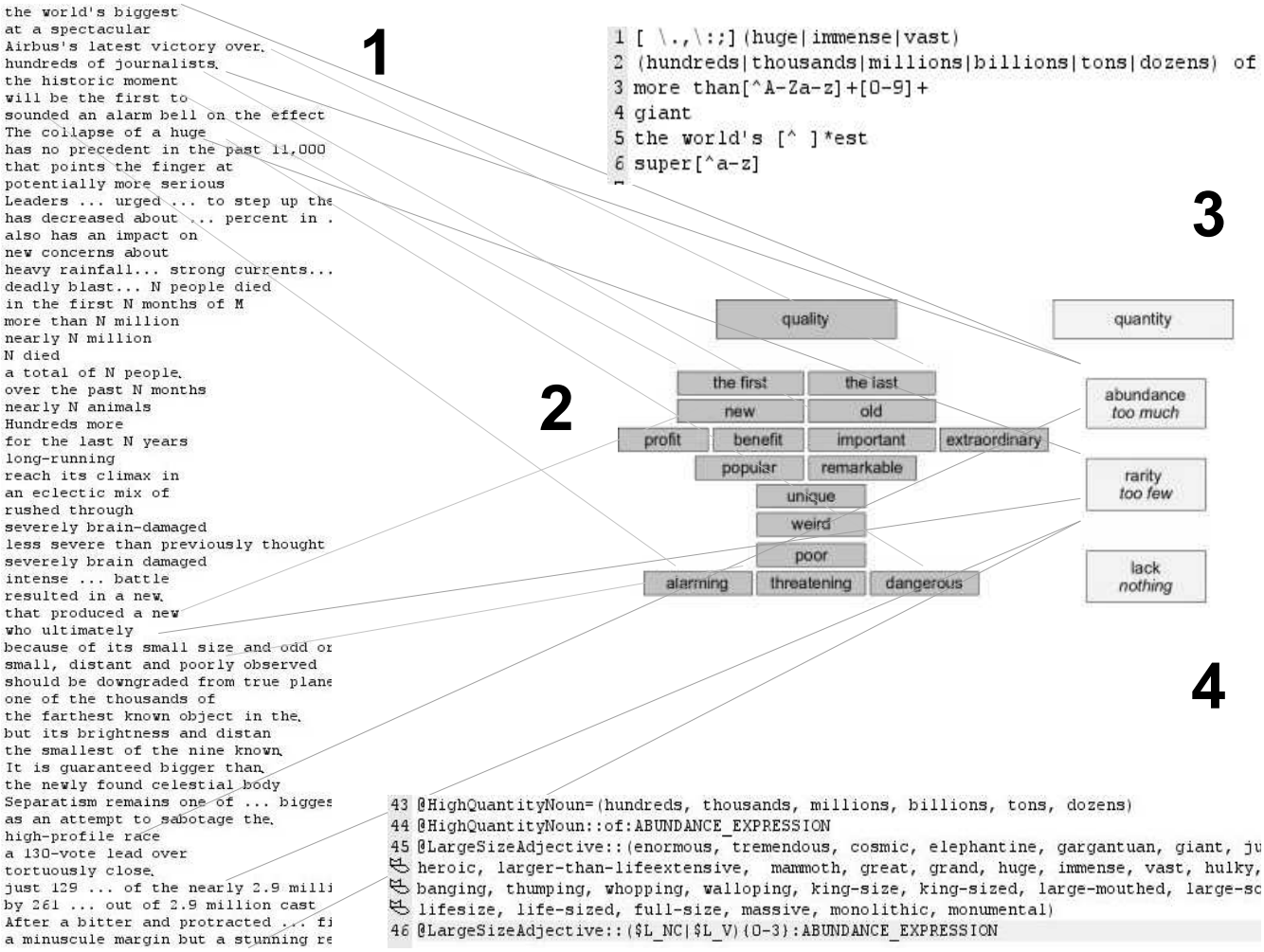

Figure 1: Contextual Exploration analysis

The resulting rules, for instance the @LargeSizeAdjective::($L_NC|$L_V)(0-3) rule of Figure 1 for identifying sentences referring to the semantic value of ABUNDANCE, are applied to the source text by the means of annotation automata during its analysis by LIMA [BdCF+ar]. After semantic analysis, system ceaList25 performs a re-ranking of the weighted sentences according to the Contextual Exploration annotation on them. Only sentences that were highly weighted by the semantic analysis are considered for re-ranking. Sentences with the highest weight and the highest number of semantic annotations are considered for extraction.

4 The Update Summarization task

Two runs were submitted to TAC 2009:

- ceaList31: Automatic summarization by sense concentration

- ceaList25: Automatic summarization by sense concentration and Contextual Exploration rules.

From the manual evaluation results, we observe that our best scores are obtained for the linguistic quality metric, which might be a result of the use of Contextual Exploration rules. Overall responsiveness and pyramid scores aren't good for set A: the use of an unlimited amount of senses punishes our set A scores, both for ceaList25 and ceaList31.

ceaList25				
	set A		**set B**	
Metric	**score**	**rank**	**score**	**rank**
Linguistic quality	5.250	17/56	5.318	15/56
Overall responsiveness	3.614	48/56	4.023	25/56
Pyramid	0.188	48/56	0.169	44/56

Table 3: ceaList25 results on manual evaluation

The automatic evaluation results of ceaList25 are slightly higher than those of ceaList31, reflecting semantic gap of our sense analysis.

ceaList25				
	set A		**set B**	
Metric	**score**	**rank**	**score**	**rank**
ROUGE-2	0.06126	48/56	0.05376	45/56
ROUGE-SU4	0.09648	48/56	0.09597	46/56
BE	0.02874	48/56	0.03059	39/56

Table 5: ceaList25 results on automatic evaluation

The results for system ceaList31 are very low. Assuming that the difference between ceaList25 and ceaList31 is statistically significant, this leads us to think that a deeper Contextual Exploration analysis might improve our results.

ceaList31				
	set A		**set B**	
Metric	**score**	**rank**	**score**	**rank**
Linguistic quality	4.227	45/56	5.318	40/56
Overall responsiveness	2.727	53/56	3.273	46/56
Pyramid	0.111	52/56	0.113	50/56

Table 7: ceaList25 results on manual evaluation

It seems that our best configuration associates a limited amount of senses plus Contextual Exploration rules. From these results, we conclude that adding limits to the semantic scope of our analysis (in this case with a topic statement) is a way of improving our results. When applying CHORAL to scientific documents, this could be done by taking document titles, or even natural language questions from the user.

ceaList31				
	set A		set B	
Metric	score	rank	score	rank
ROUGE-2	0.04850	51/56	0.04931	48/56
ROUGE-SU4	0.09005	51/56	0.09047	48/56
BE	0.02198	50/56	0.02566	46/56

Table 9: ceaList25 results on automatic evaluation

5 Evaluation of word sense-based sentence extraction

As the Update Summarization task of TAC 2009 is a complex task, it is difficult to evaluate from the results of a system to this task the contribution of each of its components. The use of word senses for sentence extraction is one specificity of the CHORAL system. Hence, we decided to evaluate more precisely the interest of such use by applying this sentence extraction mechanism to a single-document summarization task. As [Mih04], we chose the DUC 2002 corpus to perform this evaluation. This corpus is made of 567 news articles. For each article, a system to evaluate is expected to generate a 100-word summary.

Systems	ROUGE-1 average
baseline	0.4542
baseline in [Mih04]	0.4799
best DUC 2002 system	0.5011
best system in [Mih04]	0.5023
LA Times - cooc 1	0.4027
LA Times - cooc 12	0.4056
AQUAINT 2 - cooc 1 - backoff	0.4014
AQUAINT 2 - cooc 12 - backoff	0.3933
LA Times - cooc 1 - backoff	0.3986
LA Times - cooc 12 - backoff	0.3990

Table 10: Evaluation of the word sense-based sentence extraction component of the CHORAL system on DUC 2002 data

Table 10 shows the results of this evaluation. Our baseline system (first line) is the same as the baseline system of [Mih04] and generates summaries by taking the first sentences of documents until reaching the 100-word limit[2]. The results of [Mih04] for this baseline are significantly higher than ours, which must be taken into account for comparing more globally our results to the results reported in [Mih04].

Concerning the CHORAL system, three kinds of parameters related to the building of word senses were tested:

- the cohesion measure in the co-occurrence network that is computed for building the similarity matrix from which word senses are discriminated. This measure can be a first-order measure (*cooc1*) or can also take into account second-order co-occurrences (*cooc12*)[3]. The second-order cohesion measure is expected to be detect cohesion at a more semantic level and generally leads to discriminate less specific word senses;
- the corpus from which word senses are discriminated. More precisely, two different corpus were used: one middle-size corpus around 40 million words, made of the Los Angeles Times (LA Times) part of the TREC

[2]The last sentence of each generated summary is not cut if the 100-word limit is exceeded but only the first 100 words of the summary are taken into account for evaluation.

[3]The first-order cohesion measure of two words is the Pointwise Mutual Information whereas the second-order measure is the cosine measure between their vectors of direct co-occurrents.

corpus; the AQUAINT 2 corpus, which is a big-size corpus around 380 millions words. Moreover, as the evaluation documents are drawn from the AQUAINT 2 corpus, word senses discriminated from this corpus are supposed to be particularly adapted to the processing of these documents;

- the use of a backoff mechanism. When the co-occurrence subgraph of a word is not dense enough, the clustering algorithm can't discriminate any sense. This phenomenon is not rare since for the LA Times corpus, while the size of its vocabulary is equal to 30,422, senses were found for only 9,838 words. Its impact on CHORAL's results are expected to be significant as a sentence can't be extracted if no sense exists for its words. In the discriminating process of the senses of a word, a threshold is applied to its co-occurrents both on their frequency and their cohesion. By default, these two thresholds are fixed and have the same value for all words. Our backoff method consists in adapting these two thresholds according the considered word: when no sense can be discriminated with their default values, they are relaxed for having a larger number of co-occurrents. This adaptation is performed until at least one sense is produced by the clustering algorithm.

Finally, six different bases of word senses were evaluated while only the *LA Times - cooc1* (in bold into Table 10) was used for the version of CHORAL that participated to TAC 2009. The first thing to notice from Table 10 is that the results of the six versions of our system are below our baseline. This is both disappointing and not too surprising as this baseline is known to be hard to exceed. This is illustrated by the small difference in [Mih04] between the results of this baseline and those of the best system reported in this article.

The second main thing to notice is that the results of our six systems are so close to each others that they can't be considered as significantly different. Surprisingly, this means that having a larger set of word senses through our backoff mechanism doesn't improve results. More precisely, it even tends to degrade them. A more restrictive selection of the representative word senses of a document should certainly be applied to counteract this trend. The problem and its solution seem to be rather the same when word senses are built from a larger corpus such as the AQUAINT 2 corpus[4]. Finally, no global trend can be found concerning the use of a second-order cohesion measure.

6 Conclusion and further work

In this article, we have presented the application of the CHORAL system, which is initially an informative single-document summarizer, to the Update Summarization task of TAC 2009. Two versions of CHORAL were tested. The first one was a minimal adaptation of CHORAL to the Update task without taking into account topics. Unsurprisingly, its results were not good since the word senses selected for representing the conceptual content of documents were not necessarily related to the information needs expressed by topics. A second version of CHORAL was more successful by guiding the selection of the representative word senses of a document by the word senses coming from the considered topic.

The use of word senses for sentence extraction is an important aspect of the CHORAL system that was already present in our system for TAC 2008. The specificity of CHORAL in TAC 2009 is the application of Contextual Exploration rules to re-rank word sense-based selected sentences according to their role from the viewpoint of journalistic discourse. This was a first attempt to associate in CHORAL a content-based approach and a more discourse-based approach for selecting representative document sentences. Following [CFG+04], we plan to study in a more extensive way the possible interactions of these two complementary types of approaches, especially for scientific articles, which represents the main focus of the CHORAL system.

References

[BdCF+ar] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Faiza Gara, and Nasredine Semmar. Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *7th Conference on Language Resources and Evaluation (LREC 2010)*, Malta, 2010, to appear.

[4] An experiment with the AQUAINT 2 corpus but without our backoff method must be performed to confirm this fact.

[Bla08] Antoine Blais. *Résumé automatique de textes scientifiques et construction de fiches de synthèse catégorisées : Approche linguistique par annotations sémantiques et réalisation informatique*. PhD thesis, Université Paris Sorbonne, 2008.

[CFG+04] Javier Couto, Olivier Ferret, Brigitte Grau, Nicolas Hernandez, Agata Jackiewicz, Jean-Luc Minel, and Sylvie Porhiel. RÉGAL, un système pour la visualisation sélective de documents. *Revue d'Intelligence Artificielle*, 18(4):481–514, 2004.

[CG98] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[Des06] Jean-Pierre Desclés. Contextual exploration processing for discourse and automatic annotations of texts. In Geoff Sutcliffe and Randy Goebel, editors, *FLAIRS Conference*, pages 281–284. AAAI Press, 2006.

[Fer04] Olivier Ferret. Discovering word senses from a network of lexical cooccurrences. In *20^{th} International Conference on Computational Linguistics (COLING 2004)*, pages 1326–1332, Geneva, Switzerland, 2004.

[GFdC08] Jorge Garcia-Flores and Gaë de Chalendar. Syntactico-Semantic Analysis: a Hybrid Sentence Extraction Strategy for Automatic Summarization. In *MICAI 2008: Advances in Artificial Intelligence*, Atizapán, México, November 2008 2008. Springer, Lecture Notes in Artificial Intelligence.

[HL98] Eduard Hovy and Chin-Yew Lin. Automated text summarization and the Summarist system. In *Proceedings of the TIPSTER Text Program: Phase III*, pages 197–214, Baltimore, Maryland, USA, 1998.

[Mih04] Rada Mihalcea. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *42st Annual Meeting of the Association for Computational Linguistics*, pages 170–173, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[NV05] Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. Technical Report MSR-TR-2005-101, Microsoft Research, 2005.

[RHB07] Lawrence H. Reeve, Hyoil Han, and Ari D. Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management*, 43(6):1765–1776, 2007.

[Whi06] Peter R.R. White. *Evaluative semantics and ideological positioning in journalistic discourse – a new framework for analysis*, pages 37 – 69. Mediating Ideology in Text and Image: ten critical studies,. John Benjamins, Amsterdam, 2006.