MIRACLE at the Spanish WiQA Pilot: Using Named Entities and Cosine Similarity to extend Wikipedia articles

César de Pablo-Sánchez, José Luis Martínez-Fernández
Paloma Martínez
Universidad Carlos III de Madrid
{cesar.pablo,paloma.martinez}@uc3m.es
DAEDALUS S.A. - Data, Decisions and Language, S.A.
jmartinez@daedalus.es

Abstract

The WiQA pilot task explores how to select new and useful information that could be included in Wikipedia articles. Our system explores how the combination of NE and cosine similarity allow to detect new and repeated information. We have submitted two runs for the Spanish subtask wich differ in the way they select candidate sentences using the link structure in the WikipediaXML corpus. Our approach obtains results that provide at least a new snippet per topic in average. The main limitation was found in the candidate selection strategy that results in some topics being not answered or in other cases providing too much noisy candidates.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids, Novelty detection

1 Introduction

The objective of the WiQA task is to locate important and new information that could be included in a Wikipedia article. The source of the new information are other Wikipedia articles. It is assumed that an editor will judge this information as appropiate or not and she will produce the definitive article.

The WiQA pilot task has some resemblance with other IR tasks but it is unique in its combination. As passage retrieval and question answering it is expected that results are brief and focused in a topic but there is no explicit kind of information that it is sought. Like in multi-document summarization, the system must select the important information from a set of source

documents, but those could contain off-topic information. In contrast, there exist an original document that already contain relevant information and that should be completed. This setting is closer to novelty detection [5]. Finally, the selected information should account not only relevance, but importance.

The task uses the WikipediaXML [3] collection which is an XML version of Wikipedia produced for several languages. It is also a semistructured document collections that marks document format, contain tables of facts, etc. It also features a rich link structure that includes links between articles in the same language subcollection and also links between articles in different languages. For the purpose of the WiQA task, the corpus has also been annotated with additional structures, sentences and some articles have been automatically classified with basic NE classes: Person, Location and Organization.

We have develop a system for the Spanish subtask that uses the Spanish WikipediaXML subcollection. We have mainly reused components from our previous efforts at CLEF@QA[2] and have been focused mainly on what we consider the novelty detection part of the challenge. We submitted two runs to the Spanish WiQA task that use similar components but consider different ways to obtain candidate sentences.

2 System description

We have divided the system in three modules that process the information sequentally, Candidate selection, Novelty detection and Ranking. The evaluation and retrieval unit for the task is the sentence, and therefore most of the operations proceed at this level. The first module selects candidate sentences that should be included in an article. The second component detects novel sentences, or in other words, filter those sentence whose content is already present in the article or in other sentences. Finally, sentences must be ranked according to their importance to the topic.

2.1 Candidate Selection

The role of candidate selection consist on identifying sentences that are related to the article that we are completing. The article is described by a title, which for an specific language collection is unique as the Wikipedia uses the title as the primary key. When titles have an ambiguous interpretation they are complemented with information that helps to disambiguate. This is the case for topics like *Hyderabad* that could refer to the indian city or to the pakistani city (*Hyderabad (Pakistán)*) as both entries exists in Wikipedia.

Several alternatives exists to retrieve candidates as both content, title and article text, structure and the link structure seem useful for this task. In our case we have experimented only with the link structure between articles, in particular *inlinks*. We call inlinks those internal links (which use tag *collectionlink*) that refer to the query article. This approach is very precise and selects sentences that are clearly related to the query because the link is unambiguous. On the other hand, not all mentions are linked to an article, so this approach could leave important information out.

The collection has been stored using a XML database, Berkeley DBXML [1]. We use XQuery to retrieve relevant passages that have inlinks to the query article. Several structural indexes have been created to support efficient queries. We have experimented with diferent sizes for the passages, using sentences and paragraphs marked in the XML structure.

We found some practical problems with the way the collection has been converted and sentence splitting has been done. Tables and list are often recognized as one sentence and therefore lots of irrelevant information is selected. Wikipedia articles contain a lot of tables and list of facts that are not always related to the same topic. Because of that we decided to exclude this kinf od objects from candidate results.

2.2 Novelty Detection

The second module in our system is in charge of filtering the information that it is already mentioned in the original article and selecting sentences that provide new information. We have implemented novel sentence detection by a combination of two kind of information, cosine similarity between sentences and the ratio of novel an already present Named Entities (NE). A final filters candidate sentences based on three manual defined thresholds for those measures.

Cosine similarity has been proposed as a simple and effective measure to detect different levels of semantic similarity between sentences [4]. To calculate cosine similarity we index the original article in a sentence by sentence basis and compare each of the candidate sentences to them.

Original sentences as marked in the WikipediaXML corpus are processed using our language analysis toolkit composed of DAEDALUS STILUS analyzers [6] and other custom components for NERC. During development we discovered some problems with the original sentence splitting, so we decided that it could be useful to perform sentence splitting again before indexing. Stopwords are removed and simple terms withouth any other operation have been used for indexing. For indexing we have used Xapian [7] in-memory databases and their statistics to perform calculations. We obtain a pairwise cosine similariry measure with every article sentence and select the maximum value for the following operations.

Our second novelty measure is based on the presence and absence of NE in a broad sense, considering quantities and temporal expressions. We believe that the presence of these fact is probably a signal for important information in reference texts like Wikipedia. It usually signal an important relation between the focus of the article and another person, place, etc. We process the original article and produce a list of recognized NE mentions. For every candidate sentence we calculate three related measures, tha number of NE that appear in the article, the number of new NE and the ratio.

The final step consist on filtering the sentences that are not really related, too similar to the ones in the article or not important enough. The filtering is achieved by combining the previous measures and using three different thresholds. The value for this threshold have been set by manual inspection of the results in a separate and small development set. The first threshold requires a minimun number of common NE to consider the sentence. Another threshold filters sentences with a similarity. Finally, the third threshold filter those sentences in which the ratio of new NE are not significant.

2.3 Ranking

The objective of the ranking module is to present the sentences in a appropiate order to the user. We have used the information that we had available at that point. We have tried different combinations of the two measures, NE ratio of novel entities and cosine similarity. This measures have been chosen using the small development set, and again a more principled way of scoring should be considered.

3 Description of the runs

Our group submitted two runs to the WiQA subtask organized for the Spanish collection. Both runs use the same modules but in a different configuration, with different measures and parameter settings. The first run mira-IS-CN-N uses sentences that link to the article and the ratio of new entities to rank candidates. The second run mira-IP-CN-CN uses paragraphs that link to the query article to obtain initial candidates and combine cosine similarity and the ratio of new NE to rank sentences. The rank measure is proportional to the ratio of NE, and inverse proportional to the similarity. None of the runs controlled non-repetition between candidate sentences as we develop the main system in a couple of weeks.

Table 1: Evaluation results

Run	*#topics*	*#snippets*	*#I*	*#I.N*	*#I.N.NR*	*Av.Yield*	*MRR*	*P*@10
mira-IS-CN-N	50	176	96	62	54	1.080000	0.240833	0.306818
mira-IS-CN-N	67	251	127	79	69	1.029851	0.299129	0.274900
mira-IP-CN-CN	50	310	115	74	52	1.040000	0.242190	0.167742
mira-IP-CN-CN	67	431	155	95	71	1.059701	0.285359	0.164733

Table 2: Results by NE category for the official topics

Run	*#category*	*#topics*	*Av.Yield*	*MRR*	*P*@10
mira-IS-CN-N	P	17	0.764706	0.166667	0.302326
mira-IP-CN-CN	P	17	1.470588	0.309384	0.223214
mira-IS-CN-N	L	17	1.117647	0.247549	0.287879
mira-IP-CN-CN	L	17	0.470588	0.150000	0.076923
mira-IS-CN-N	O	16	1.375000	0.312500	0.328358
mira-IP-CN-CN	O	16	1.187500	0.268750	0.202128

4 Results

The Spanish subtask consisted in a total of 67 topics that were developed by participant teams. The 50 first topic followed the guidelines for topic creation as close as possible. Only topics tagged as PERSON, LOCATION or ORGANIZATION in the collection were considered as official ones. These topics have different article length distribution, from stubs (very short articles) to longer articles. The other 17 topics are considered additional. Some of them were left over randomly from the official set and others do not strictly follow the guidelines.

Table 1 summarizes the results obtained by the two runs and presents the runs in the official results and over all the developed topics. There are significant differences between the two runs, especially in the number of snippets returned by each run. As we already expected, run mira-IP-CN-CN, that consider more candidate sentences, returns more results. If we consider that the maximum number of snippets a system could return 670, both runs have low number of results. This is the consequence of our method for candidate selection. This is also the reason for some topics having no results. In contrast, the system is able to provide at least one interesting snippet in average per topic.

In Table 2, we analyze the performance of the runs regarding the different class of topics from the official set. Topics where classified as Person (P), Location (L) and Organization (O). Significant differences are appreciated in each of the classes for the two runs. For P topics the run based on paragraphs performs much better, while the results are completely different for L topics. We believe that the main reason is that candidate selection in run mira-IP-CN-CN works different for those topics. When persons are mentioned in a paragraph it is probable that the rest of the paragraph also mentioned related facts. In contrast, lots of the references for locations are places where someone was born or died. The rest of the article is probably not related to the location article we would like to extend.

5 Conclusions and Future Work

Our system obtains moderate good results for the task, but a larger yield of snippets is needed for a tool to be useful to complete the proposed task. We have considered a rather precise approach to the way we select candidates initially which is the main reason for the low average yield. We expect to focus on this module to include other sources and forms of selecting candidate sentences that are more recall oriented.

One of the most interesting things about WiQA is that it seems an interesting task for a multilingual setting. Combining multilingual information would definitely help creating Wikipedia topics, as far as the editor understand several languages. We would like to explore this setting in future editions. For example, we have used some shallow language analysis in the NE detection module, but we believe that the mark-up structure of the corpus could be used for a similar purpose. In this way, the system could be extended to other languages with little effort.

Finally, more investigation is needed to estimate importance models that reflect the importance of relations between entities described by the articles in real world and not just by textual snippets.

References

[1] Sleepycat berkeley db xml 2.2. On line http://www.sleepycat.com/products/bdbxml.html, July 2006. last visit.

[2] de Pablo-Sanchez C. et al. Miracle's 2005 approach to cross-lingual question answering. In *Working Notes for the CLEF 2005 Workshop. Vienna,Austria*, 2005.

[3] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[4] Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524, New York, NY, USA, 2005. ACM Press.

[5] Ian Soboroff. Overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[6] Stilus website. On line http://www.daedalus.es, July 2006.

[7] Xapian: an open source probabilistic information retrieval library. On line http://www.xapian.org, July 2006. last visit.