Wikipedia entity retrieval for Dutch and Spanish

Gosse Bouma and Sergio Duarte
Information Science, University of Groningen
g.bouma@rug.nl, sergio.duarte@gmail.com

Abstract

We developed two systems (for Dutch and Spanish) for the GikiCLEF task, in which Wikipedia pages have to be found that match a description in natural language. We concentrated on linguistic analysis of the query, for mapping the question onto the most relevant Wikipedia categories, and for extracting additional constraints that matching pages have to satisfy. In addition, for Spanish we experimented with query expansion for improved recall of the IR process. In both the Dutch and Spanish system we tried to incorporate additional knowledge sources (WordNet, Yago, DbPedia) for better question analysis and retrieval results. The Dutch system obtained a GikiCLEF score of 2.5 (7th overall and 7th for Dutch). The Spanish system was still under development at the time of the official evaluation, and performed poorly. We show that the completed system would have performed well at the 2009 task.

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 Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Entity Ranking, Wikipedia, Linguistic Analysis, Dutch, Spanish

1 Introduction

The GikiCLEF entity retrieval task offers a number of challenges for traditional IR systems. Relevant Wikipedia pages for queries such as *List the basic elements of the cassata* are not easily found on the basis of keyword index. One reason is that apart from relevance, it is also important to return only pages that satisfy a given category (i.e. not the page for cassata itself). A second reason is that the additional constraints are typically so that they require more than simple keyword matching. Numerical constraints (*more than 1 million inhabitants*) and non-trivial geographical constraints (*born in the Bohemian Forest*) require reasoning over numbers and geographical locations.

For this reason, we were especially interested in developing a system that incorporates knowledge from external sources (i.e. WordNet) and which uses knowledge harvested from Wikipedia itself (i.e. property-value tuples extracted from infoboxes).

The system we developed for Dutch uses a Dutch Wordnet as part of the module that determines appropriate categories for pages that can be retrieved. Furthermore, it has access to

relational information harvested from infoboxes in Dutch Wikipedia and expanded with information extracted from infoboxes found in English Wikipedia. The Spanish system uses, among others, cross-language links and English Wordnet for query expansion. For Dutch we were able to use an existing full parsing system for syntactic analysis of the question. For Spanish, an existing POS tagger was combined with a NE tagger trained on CONLL data.

It should be noted that the Spanish system was still under development when results had to be submitted. As a consequence, we were not able to submit a single run combining the results for Dutch and Spanish. We describe the Dutch and Spanish system in sections 2 and 3, respectively. Some suggestions for future work are given in section 4.

2 Dutch

The Dutch system for the GikiCLEF task consists of a module for query analysis, which predicts appropprate categories for a given query, and which tries to identify additional constraints that returned pages must satisfy, and an IR component which returns the most relevant pages based on an index that was developed for the QA task of CLEF 2008. We use a simple ranking scheme that prefers pages that satisfy the categorical constraints, the additional constraints (if any are found), and finally ranks all pages that satisfy these conditions on the basis of the IR score. At most 15 pages are returned.

The goal of query analysis is to identify words in the input that can be used to predict the most appropriate Wikipedia category for a query (e.g. *African capital*), and to identify additional constraints that matching pages have to satisfy (e.g. *more than one million inhabitants*). We use a syntactic parser for Dutch [10] to parse the query. From the parse result, we extract root forms, part-of-speech labels, and dependency relations.

Below, we first describe how category labels are predicted, and next how additional constraints are identified. We conclude with a discussion of the results obtained for Dutch, including error analysis.

2.1 Predicting Category Labels

Wikipedia pages are typically classified into one or more categories such as *Dutch author*. Queries often contain phrases that are similar to the category labels used to classify Wikipedia pages (i.e. *writers from the Netherlands*. However, as queries rarely contain a phrase that literally matches a Wikipedia category, we try to find the optimal Wikipedia category matching a query.

To this end, we parsed all category labels used in Dutch Wikipedia using the Alpino parser, and stored the head noun, its stem, and its root form (for compounds). Furthermore, for each category a list of content words modifying the head noun is stored. Finally, head nouns are linked to a corresponding word in a Dutch Wordnet (Cornetto[1], [11]). An identity-link is established if there are no modifiers, and the head noun is found in the wordnet. An isa-link is established if the root form can be found in the wordnet. For instance the category labels *Attractiepark* (*Amusement park*), *Berg in Chili* (*Mountain in Chile*) and *Amerikaans stripauteur* (*American comic book author*) are stored as

⟨Attractiepark, Atrractiepark, atrractie_park, park, [ ], ident, dn24615⟩
⟨Berg in Chili, Berg, berg, berg, [Chili], isa, dn32731⟩
⟨Amerikaans stripauteur, stripauteur, strip_auteur, auteur, [Amerikaans], isa, dn25105⟩

Given a query parsed by Alpino, we now determine the most appropriate Wikipedia category labels as follows:

1. Potential head nouns and their modifiers are identified. Head nouns are linked to all wordnet ids for this noun,

[1] http://www2.let.vu.nl/oz/cltl/cornetto/

2. Wikipedia categories are retrieved whose wordnet id matches with one of the ids found for the given headnoun,

3. A score for the retrieved category is computed, based on the overlap between modifiers present in the query and the category label,

4. The highest scoring category labels are selected.

Potential head nouns are all nouns in the query which are not themselves part of a phrase that is a modifier to a noun. Note that we do not use the isa-relationships between categories in Wikipedia. This was motivated by [9], who observe that the Wikipedia category system contains many non-taxonomic links. As an alternative, they propose to link categories for individual Wikipedia pages to (English) WordNet word senses, and to use the WordNet hypernym relations as an alternative for category isa-relations. We created a similar resource for Dutch [3].

Given a query such as *Welke Nederlandse violisten...* (*which Dutch violin players...*), the noun *violisten* (with stem *violist*) is identified as a potential head. Matching categories are *Duits violist, Amerikaans jazzviolist, Amerikaans altviolist, Nederlands violist, Nederlands jazzviolist, etc.* The latter two are selected as the most relevant categories, as they contain a modifier *Nederlands* that also occurs as a modifier in the query.

The link to the wordnet allows us to search using synonym and hypernym relations. That is, if a query contains the word *schrijver* (*writer*), we might still consider *Amerikaans stripauteur* as a matching category label, as *schrijver* and *auteur* (*author*) are synonyms. Given the query *musikant* (*musician*), we also find *violist*. A problem with this approach is that the linking of Wikipedia labels to wordnet senses requires word sense disambiguation, as most nouns have multiple meanings in wordnet. We solved this problem for Dutch by choosing the predominant word sense on the basis of distributional similarity data obtained from a large Dutch corpus, following the idea of [8]. This method is not perfect, however, and this has a negative effect on the selection of Wikipedia categories. All categories for *sharks* and *snakes*, for instance, are linked to wordnet senses denoting negative characterizations of female persons. As a consequence, queries for *vrouw* (*women*) may match with Wikipedia categories for *snakes* and *sharks*.

2.2 Templates

Apart from a categorical constraint, queries often impose numerical or geographical constraints on what constitutes a valid answer: *geboren in het Bohemer woud, geboren in Alaska* (*born in Bohemia, born in Alaska*), *met twee of drie Michelinsterren, met meer dan 10.000 studenten* (*with two or three Michelin stars, with more than 10,000 students*). Such constraints are not easily checked using a plain IR engine, as a page may contain both the words *born* and *Alaska*, without containing the information that Alaska was the birthplace of the entity described by the page. Numerical expressions like *more than 10,000 students* can be satisfied by pages which do not contain the number 10,000.

Many Wikipedia pages contain a so-called infobox, expressing the most relevant information for a given entity. For instance, web pages for universities usually mention the number of students in the infobox (see figure 1).

We stored all information in all infoboxes in Dutch Wikipedia and stored the result as relation tuples $\langle$*Page, Attribute, Value*$\rangle$. As English Wikipedia typically contains more elaborate infoboxes for a given entity than the corresponding Dutch page, we also automatically expanded the set of relation tuples with tuples harvested from English Wikipedia. Attribute names were automatically translated into Dutch as well (see [2] for details).

For queries such as *Nederlandse universiteiten met meer dan 10.000 studenten* (*Dutch universities with more than 10,000 students*), question analysis finds the constraint `more_than(students, 10000)`. For potentially matching pages P, we can now check whether there exists a tuple $\langle P$, students, $V\rangle$ where $V > 10000$.

Figure 1: Infobox for University of Groningen

2.3 Results and Error Analysis

Our system returned 638 answers, 36 were correct. Thus we obtained a precision of 0.05 and a GikiCLEF score of 2.4 (7th). For Dutch only, we returned 502 answers, of which 22 were correct (0.04 precision, 0.9 GikiCLEF score).

Identification of appropriate Wikipedia categories on the basis of the queries turned out to be hard. In only 17 out of 50 questions, one or more categories were identified that were considered appropriate for identifying the correct pages. In 9 cases, no category could be found. This happened for instance for queries containing uncommon (compound) nouns (*basiselementen, Formule-1-rijders, talentenjachtwinnaars* (*basic elements, Formula 1 drivers, talent show winners*), but also for common nouns such as *plaatsen, bergtoppen* en *skioorden* (*mountain tops, ski resorts*, and *places*), which are not used as category labels in Wikipedia, and which also could not be linked to categories using synonyms. Errors are caused by nouns such as *landen* and *rivieren* (*countries* and *rivers*), which correspond to highly general Wikipedia categories, for which several tens of more specific subcategories exist, covering a large number of Wikipedia pages. Such categories are not very selective, and easily lead the system to prefer results from a subcategory unrelated to the input question. Sometimes, the wrong noun in the query was selected as the head noun of the category.

Additional numerical and geographical constraints were only correctly identified for two questions. This module was not effective during the experiments.

Finally, we noted that the IR module performed poorly in terms of recall. This means that many of the pages that fall within one of the categories identified for a question were not in the set of pages retrieved by means of IR. As a consequence, such pages cannot be ranked on the basis of their IR score.

3 Spanish

The Spanish system was developed to extract entities (wikipages) from the Wikipedia Spanish Collection according an input topic given in natural language. The input topics are parsed to obtain noun phrases (NPs) representing two types of information: The target named entity (NE) type and the restrictions or constraints on the desired NEs.

The target NE type is defined with a set of Wikipedia categories and a set of Yago and DBpedia classes. These items are obtained using a text search on an IR engine (Lucene). We built independent indexes containing the Wikipedia categories titles and the list of Yago and DBpedia types. Although this matching can be also done using a simple look-up table, we think that this approach is more convenient because the query expansion methods described in section 3.3 can be applied in a transparent fashion.

A candidate entity set is constructed from the wikipages belonging to these Wikipedia categories. Given that the wikipages of the same category can have different NEs types, we clean the set by ignoring the entities that do not correspond to the target YAGO/DBPedia classes.

To evaluate the restrictions on the NEs, we first map the NPs that specify constraints to Wikipedia categories as it was done with the mapping of the NE type. Additionally, we obtain the wikipages associated with the NEs mentioned in these NPs. We construct a set of wikipages for each NP adding the members of the categories found and the wikipages of the NE. The phrases that cannot be mapped to categories or do not mention any NE are matched in the text content or infobox of the set of wikipages constructed.

This matching is done using the IR engine, for this purpose a temporal index is created with the pages of these sets. Pages scored below a threshold are deleted. In the final step, we create a second set of candidate entities by including the entities that point to or are pointed from the entities of the restriction sets, as it is done in the *WikipediaListQA@wlv* system [5]. This set is also filtered to include only those entities that match the target NE type. We return as result the intersection of the two candidate entity sets (the one created with the first target NE type noun phrases and the one created with the restriction noun phrases). We return only one set in case the other one is empty. In the following sections we describe in further details the main features used in the system.

3.1 Shallow parsing of the input topic

The input topics are parsed by first tagging the tokens with their POS and then extracting the NPs. We observed that the NE type is commonly specified in the first NP encountered in the user's topic. This is the case for all the GikiCLEF topics in Spanish. Further information concerning these NEs (restrictions) is specified in the NPs functioning as object of the sentence or post-modifiers of the main NP. These two NPs are frequently separated by relative pronouns such as *that* or *in which*. In more complex sentences the phrases are separated by verb elements. Thus, first the verb phrases or relative pronouns are detected to split the NPs that refer to the NE type and the ones that refer to the restrictions. This is done by matching the tokens with the relative pronoun POS tags and with a list of common particles used to introduce relative clauses such as: *en los cuales* (*in which*), *en la que* (*inside which*), *por el que* (*along which*), *a la que* (*whom*) and so on. Similarly, verb phrases are extracted by matching the rule *(Aux Verb)* (Verb) (Adv)**. This process leads to a list of NPs in which the first element defines the NE type. Each one of these NPs is further subdivided to allowing only prepositional phrases as post-modifiers. This further splitting is carried out because the motivation is to match Wikipedia Categories to these NPs and the categories are commonly described by short NPs.

Applying this method on the topic *Nombre los lugares de Italia que haya visitado Ernest Hemingway a lo largo de su vida (List the Italian places where Ernest Hemingway visited during his life)* lead to the following phrases: {*lugares de Italia*}, {*Ernest Hemingway*}, {*a lo largo de su vida*}.

The POS tagging is performed using the OpenNLP POS Spanish tagger which employs a maximum entropy model to predict the POS of each word. We encountered some inaccuracies using

this tagger on the GikiCLEF topics and other simple sentences, making it hard to identify phrases using methods based on the POS. For this reason a procedure was designed to tune the results and increase the accuracy of the tagger. First, missing nouns and numbers are detected using the Stanford NER parser[7][2]. Misclassified tags belonging to the closed POS classes such as determiners, conjunctions and prepositions are corrected by looking up a table with a comprehensive list of words with these POS classes. This list is extracted from the EAGLES tag definition of the Technical University of Catalonia (UPC) and from Wikipedia. Currently this list contains around 420 items. Further inconsistencies are detected by checking contextual and lexical rules to find sequences of POS that are unfeasible in the Spanish grammar. Similar rules are used to detect the most likely correct tag.

3.2 Yago and DBpedia resources

Yago is a semantic knowledge base extracted from Wikipedia and WordNet [9]. Yago contains more than 2 million entities and 20 million facts about these entities which are available to the public under the GNU license. Particularly, we are utilizing the Yago type definition which assigns a set of types to each Wikipage in the English collection. This ontological classification is based on the conceptual categories of Wikipedia and the WordNet synsets[9]. We found 4930 different types in the data which were translated to Spanish by using the cross-lingual dictionary and then cleaning and completing the results by hand. The cross-lingual dictionary extracted contains 112.099 entries.

Conversely, DBpedia is a community effort to extract information from Wikipedia and interlink this information with other knowledge bases available on the Web [1]. We particularly employed the types defined in the DBpedia Ontology and the set of types assigned to the wikipages in the English collection. This ontology was hand-generated from the Wikipedia infoboxes and it contains 170 classes. Although the hand-generated mapping does not cover all the possible infoboxes and properties present in the Wikipedia collection, the most frequent infoboxes are included and normalized. Since the types are defined only for English, we translate the 170 classes to Spanish manually.

The filtering of entities in a set is performed applying the following steps for each entity:

1. Obtain the English name of the Spanish Wikipage using the cross-lingual links
2. Fetch the types of the English Wikipage in the Yago/DBpedia data
3. Intersect the types fetched in step 2 with the target NE types obtained from the input question. If the size of the intersection is below n[3], discard the Spanish Wikipage under evaluation.

3.3 Query processing and expansion

The queries are constructed including the nouns, adjectives and adverbs of the NPs. These tokens are expanded using three query expansion methods:

3.3.1 Ontological Expansion

In this expansion method we employed the DBpedia ontology to include in the query all the types under the hierarchy of the DBpedia types found in the input NPs. In this ontology the classes and subclasses are related with the hypernymy semantic relation. For instance if the user requires information about *german artists*, the algorithm expand the terms to include *german writers, comedians, actors, etc.*, since *writer, comedian* and *actor* are types under the type *artist* as it is shown in figure 2.

[2]This NER was trained using the Spanish data provided in the shared task of CoNLL 2002 [4]

[3]In our system we set n to 2

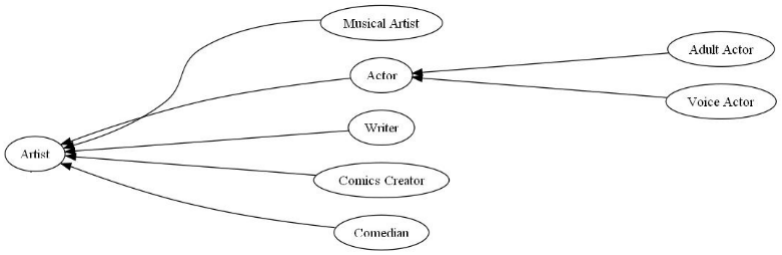

Figure 2: Fragment of DBPedia Ontology showing the subclasses of the type Artist

3.3.2 WordNet Expansion

WordNet [6] is used to include the synonyms of the nouns in the input sentence. We made use of the cross-lingual links of Wikipedia to translate the target Spanish word to English. Then we extracted the synonyms of the English nouns and these are translated back to Spanish using again the cross-lingual links. Although the coverage of this method is low given that we are only using the dictionary built from the cross-lingual links, we found that this method led to a slight increase of the performance of the system.

3.3.3 Redirect Links Expansion

We found that the redirect links can be exploited to normalize and nominalize entity names. These links are used in Wikipedia to let the user refer to an entity using alternative names and word forms such as: pseudonyms (e.g. *Samuel Langhorne Clemens* redirects to *Marc Twain*), abbreviations, common misspelling forms (e.g. *Condoleeza Rice* redirects to *Condoleezza Rice*), alternative spellings (e.g. *colour* redirects to *color*) and other adjectival forms (e.g. *Peruvian* redirects to *Peru*). The dictionary of redirect links extracted from the GikiCLEF Wikipedia collection contains 113.790 Spanish entries. Nonetheless missing pairs of country-demonyms were added to the dictionary since this type of information is frequently used in the GikiCLEF topics.

3.4 Results

Unfortunately by the official submission deadline the system was still on its middle development stages, for this reason it was only possible to submit a result list with a system implementing very few of the features described above which lead to poor results. Nonetheless, we collected all the results submitted by the 17 participants to build a list with the correct answers for the 50 topics included in this track edition. We utilized this list to evaluate the performance of our final system and evaluate the impact of the ontological resources and the query expansion methods. This list contains 105 correct answers in Spanish.

We evaluated our system using all the query expansion procedures, the category expansion and the ontological resources and we obtained a GikiCLEF score of 4.08 (0.168 precision and 0.238 recall). These results are highly promising given that the system would rank in the fourth position among the 17 participants of the task (considering only Spanish) and because we only processed the Spanish Wikipedia collection, which is significantly less developed and completed than the English collection.

A baseline system was set to evaluate the performance gain obtained by the used of the DBpedia resources, the query expansion methods and the category expansion. This system excludes the used of all these features. The results are summarized in table 1. From these results we observed that the query expansion methods and the Yago/DBpedia type filtering provide the greatest overall performance gain in the system.

Similarly, we evaluate the performance gain obtained by each query expansion method. As baseline we include the Yago/DBpedia filtering because the query expansions techniques are also

Setting	GikiCLEF score	Precision	Recall
Baseline (B)	0.73	0.06	0.11
B. +Yago/DBpedia	1.42	0.80	0.13
B. +Query Expansion	1.83	0.96	0.18
B. +Category Expansion	1.18	0.08	0.13

Table 1: Performance gain for some of the features of the system

Setting	GikiCLEF score	Precision	Recall
Base + Yago/DBpedia (B-YD)	1.42	0.80	0.13
B-YD. + Semantic Exp.	1.52	0.90	0.14
B- YD. +Redirect L. Exp.	1.92	0.11	0.16
B-YD. +WordNet Exp.	1.56	0.94	0.14

Table 2: Performance gain for the query expansion methods

used in the matching of types. Results are summarized in table 2. We found that the query expansion based on the Wikipedia redirect links contributes the most to increase the overall performance of the system. Redirect links are commonly used to obtain the country that corresponds to demonyms or adjectival forms expressed in the topics. This situation appears in 46% of the Spanish GikiCLEF topics.

4 Conclusions

We have developed two systems for Wikipedia entity retrieval based on linguistic analysis of the query, query expansion, and incorporation of knowledge harvested from Wikipedia itself, and from external knowledge sources. Both systems have their own strenghts and weaknesses. Linguistic analysis is smoother for Dutch, given the fact that we could use a full parser in combination with a Dutch Wordnet. The Spanish system uses more shallow syntactic analysis and consults Wordnet through cross-language links. On the other hand, the IR component of the Spanish system is much more sophisticated and more targeted towards the task of entity retrieval.

An obvious direction for future work is to develop an integrated system that employs sophisticated IR for both languages, and which also remedies some of the weaker aspects of the linguistic analysis for Spanish.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2008.

[2] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual Alignment and Completion of Wikipedia Templates. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)*, pages 61–69, Boulder, Colorado, 2009. Association for Computational Linguistics.

[3] Gosse Bouma. Linking Dutch Wikipedia Categories to EuroWordNet. In *Proceedings of the 19th Computational Linguistics in the Netherlands meeting (CLIN 19)*. Groningen, the Netherlands, 2009.

[4] X. Carreras. Resources on named entity recognition and classification, 2002.

[5] D. Santos et al. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. In *Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[6] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT, Cambridge, 1998.

[7] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, 2005.

[8] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590, 2007.

[9] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM Press.

[10] Gertjan van Noord. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. 2006.

[11] P. Vossen, I. Maks, R. Segers, and H. van der Vliet. Integrating lexical units, synsets, and ontology in the cornetto database. In *Proceedings of LREC-2008*, Marrakech, Morocco, 2008.