Wikipedia-based approach for linking ontology concepts to their realisations in text

Giulio Paci, Giorgio Pedrazzi, Roberta Turra

CINECA - Consorzio Interuniversitario
via Magnanelli 6/3, 40033 Casalecchio di Reno (BO), Italy
{mining, g.paci, g.pedrazzi, r.turra}@cineca.it

Abstract

A novel method to automatically associate ontological concepts to their realisations in texts is presented. The method has been developed in the context of the Papyrus project to annotate texts and audio transcripts with a set of relevant concepts from the Papyrus News Ontology. To avoid strong dependency on a specific ontology, the annotation process starts by performing a Wikipedia-based annotation of news items: the most relevant keywords are detected and the Wikipedia pages that best describe their actual meaning are identified. In a later step this annotation is translated into an Ontology-based one: keywords are connected to the most appropriate ontology classes on the basis of a relatedness measure that relies on Wikipedia knowledge. Wikipedia-annotation provides a domain independent abstraction layer that simplify the adaptation of the approach to other domains and ontologies. Evaluation has been performed on a set of manually annotated news, resulting in 58% F_1 score for relevant Wikipedia pages and 64% for relevant ontology concepts identification.

1. Introduction

The paper briefly describes a system that automatically identifies realisations of concepts in texts. This work has been developed in the EU-funded Papyrus project *Cultural and Historical Digital Libraries Dynamically Mined from News Archives* (`http://www.ict-papyrus.eu/`). The project aims to create a dynamic digital library which understands user queries in the context of a specific discipline (e.g.: history), look for content in a domain alien to that discipline (e.g.: news) and return the results presented in a way useful and comprehensive to the user. The project plans to achieve this by modelling the two disciplines with two ontologies and perform mapping between them. The content of the library has to be analysed and relevant concepts have to be identified and connected to the proper ontology classes. News items (video and text), specific to the renewable energy and biotechnology domains, have been provided for the prototype by Agence France Press (AFP) and Deutsche Welle (DW) and have been annotated using the proposed system.

The system performs a Wikipedia-based annotation of news items followed by the translation of this annotation into an Ontology-based one. These two levels of annotation are both useful to summarise documents and can be used to provide useful hints about the content. Ontology-based annotation connects documents to the ontology, while Wikipedia-based annotation is used to enrich them with semantic metadata for indexing purpose and to support the ontologists with either suggestions about possible missing concepts or integrations of concepts definitions with Wikipedia information.

One of the purposes of the proposed annotation method is to avoid strong dependency on a specific domain ontology. Wikipedia-annotation provides a domain independent abstraction layer that simplify the adaptation of the approach to other domains and ontologies.

After exploring related works, in section 3. a brief overview of the underlying Wikipedia analysis is provided. In section 4. a general overview of the automatic annotation system is presented, followed by more in-depth description of all the tasks involved. In section 5., the annotation set used as gold standard for the evaluation is described. Finally, section 6. provides the system evaluation results.

2. Related work

In (Milne et al., 2006; Medelyan and Milne, 2008) it's shown how the classic thesaurus structure of terms can be mined automatically from Wikipedia. It's also shown that, in a comparison with a professional thesaurus for agriculture, Wikipedia contains a substantial proportion of its concepts and semantic relations. Furthermore Wikipedia has an impressive coverage of contemporary documents in that domain. This justifies an attempt to use Wikipedia information to perform specific domain annotations and to semi-automatically improve an ontology of such domain.

Several studies have been undertaken on automatic text annotation based on Wikipedia knowledge (Milne and Witten, 2008b; Mihalcea and Csomai, 2007), with the aim to mimic Wikipedia users' behaviour. Our work is largely based on the Wikipedia miner toolkit (Milne, 2009) that provides easy access to Wikipedia and statistics about it that can be used to compute concepts similarity measures and to perform word sense disambiguation. The Wikipedia-based annotation is very similar to the one described in (Milne and Witten, 2008b), although we used a different relevance criterion, that exploits domain information, to select relevant concepts.

In (Ruiz-Casado et al., 2005) a method is described to link Wikipedia articles with concepts in a lexical semantic network (WordNet) by the use of a similarity measure between synset's definitions and Wikipedia page's short descriptions. In (Reiter et al., 2008) for each class in the ontology, the most appropriate Wikipedia articles are associated to it using several variants of matching a set of domain terms against the articles. In (Medelyan and Milne, 2008) a similar mapping is proposed, based on the Wikipedia Miner semantic relatedness measure described in (Milne and Witten, 2008a).

To map Wikipedia-based annotation into an ontology-based one, we used techniques similar to the ones cited above, while taking advantage of document content when the same Wikipedia page refers to more than one ontology concept or vice-versa.

3. Wikipedia description

Wikipedia is a free, web-based, collaborative, multilingual encyclopedia. Wikipedia content is presented on pages:

- articles: contain encyclopaedic information. Each article should describe a single concept and should begin with a brief overview of the topic, for each concept there should be only one article;
- redirects: redirect users to another page. They encode pluralisms, technical terms, common misspellings and other variants;
- disambiguation pages: contain a list of articles corresponding to different meaning of the same word;
- categories: are nodes for hierarchical organisation of articles.

3.1. Wikipedia Miner

Wikipedia Miner (Milne, 2009) is a toolkit that uses Wikipedia as a linguistic resource and provides access to its structure and content. Moreover it provides useful statistics about page anchors (text used by Wikipedia authors when linking a page) and links.
The toolkit provides an advanced search based on anchors (anchor-search) that can be used to retrieve a list of all the Wikipedia articles that are referred using the same anchor (senses of the anchor), described in section 4.1..
Among the statistics provided are:

1. commonness of an anchor sense, $\frac{N_s}{N_a}$;
2. link probability of an anchor text, $\frac{N_a}{N_t}$.

where N_s is the number of times the anchor is used to link to this sense, N_a is the number of the times the anchor text is used as an anchor, N_t is the number of times the anchor text is used in Wikipedia.
Wikipedia Miner also defines

- a relatedness measure (Milne and Witten, 2008a) between two senses a and b (term relatedness);

$$R = \frac{\log\left(\max(|A|, |B|)\right) - \log\left(|A \cap B|\right)}{\log\left(|W|\right) - \log\left(\min(|A|, |B|)\right)} \qquad (1)$$

  where A, B and W are the set of the links going in and out pages a, b and all the pages in Wikipedia;
- relatedness measure between a sense and a set of articles (context relatedness, see section 4.1.): the weighted average of the relatedness between the sense and each context article.

Finally Wikipedia Miner provides

- word sense disambiguation based on these measures (see section 4.1.);
- a tool to annotate documents with links that a Wikipedia author would provide if the documents were Wikipedia pages.

4. Procedure

The annotation system has been developed for the Papyrus project to annotate textual news items and automatically transcribed videos, with these main goals:

- summarise documents with the most important keywords;
- enrich documents with additional information (semantic metadata) for indexing purpose;
- connect identified keywords with concepts in a specific domain ontology;
- suggest possible missing or duplicated ontology concepts.

The system is composed by a Wikipedia-based annotator, that performs generic domain annotation, followed by a component that translates this annotation into a domain specific, ontology-based, one.
The automatic Wikipedia-based annotation procedure is similar to the one described in (Milne and Witten, 2008b), although we're using a shallow parser, based on treetagger (Schmid, 1994), to identify possible relevant concepts and a different relevance criterion to filter them. The annotation procedure can be divided into the following logical steps:

1. candidate keywords identification: the text is processed with a part-of-speech tagger, and nominal phrases are identified;
2. anchor-search of all the candidate keywords: the search result is used as the disambiguation context (see section 4.1.);
3. keyword disambiguation: the proper Wikipedia page is associated to each keyword (see section 4.1.);
4. relevant keyword selection: relatedness measures are computed with respect to the document and to the ontology domain (see section 4.2.).

The Wikipedia-based annotation is translated into an ontology-based one by converting each *keyword-Wikipedia page* connection into a *keyword-Ontology class* connection. This is achieved using a pre-computed hash table (described in section 4.3.) that uses both the keyword in the text and the associated Wikipedia page as keys.

4.1. Candidate keywords identification and disambiguation

The Wikipedia-based annotation starts with a shallow parsing procedure based on the TreeTagger chunker output. Nominal phrases (adjectives and nouns, without articles) are extracted and used as candidate keywords. For each candidate keyword a Wikipedia anchor-search is performed, to detect pages that can explain the actual meaning of the keyword. The anchor search implemented in Wikipedia Miner is performed by searching the keyword text

among all the anchors created by Wikipedia users and retrieving all the pages linked by those anchors. Search results can be improved by applying preprocessing to both keyword texts and anchors. In our system preprocessing has been applied by means of the Snowball stemmers and stop-word lists collection (Porter, 2001) that we integrated in Wikipedia Miner.

The disambiguation step takes care of selecting the page that better describes the actual meaning of the keyword in the document among those retrieved by the anchor-search. The disambiguation facility is provided by Wikipedia Miner using a machine learning approach (Milne and Witten, 2008b). The commonness of senses (see section 3.1.), their relatedness to the surrounding context (context relatedness) and the context quality are used as features to train a classifier. Links found within Wikipedia articles are used as training set: for each article anchors destinations are considered as positive examples, while all other possible senses of that anchor are considered as negative ones.

Computing the context relatedness requires the definition of a context which poses a cyclic problem because these terms may also be ambiguous. To solve this issue unambiguous link are used to disambiguate ambiguous ones and the final result is used as a disambiguation context.

Finally, the context relatedness of a sense is defined as the weighted average of its relatedness to each Wikipedia page in the context. The weight applied to each term of the context is the average between its link probability and the average of its relatedness to all the other context terms. The link probability gives an indication of a term usual semantic significance, while the average relatedness gives an indication of its relatedness to the central thread of the context (Milne and Witten, 2008b).

The context quality is given by the sum of the weights that were previously assigned to each context term. This takes into account the number of terms involved, the extent they relate to each other, and how often they are used as Wikipedia links.

The three features are used to train a classifier that produces a disambiguation probability for each sense.

In our system we selected only the sense with the highest probability, so that only one page is associated to each keyword.

4.2. Keywords selection

The context relatedness of a Wikipedia page, that up to now has been evaluated with respect to the other concepts appearing in the text under analysis, can be evaluated against a generic set of Wikipedia pages. This measure is really useful for selecting relevant concepts when the document internal context is unreliable as is the case in transcripts resulting from automatic speech recognition systems (Paci et al., 2010). A domain specific context can be easily created by collecting terms peculiar to that domain and disambiguating them. This context can be used to evaluate the relatedness of a Wikipedia page with the specific domain (domain relatedness). In Papyrus domain contexts for the renewable energy and biotechnology domains have been created by collecting documents in those domains, extracting chunks and filtering them according to a manually defined threshold on their TF-IDF.

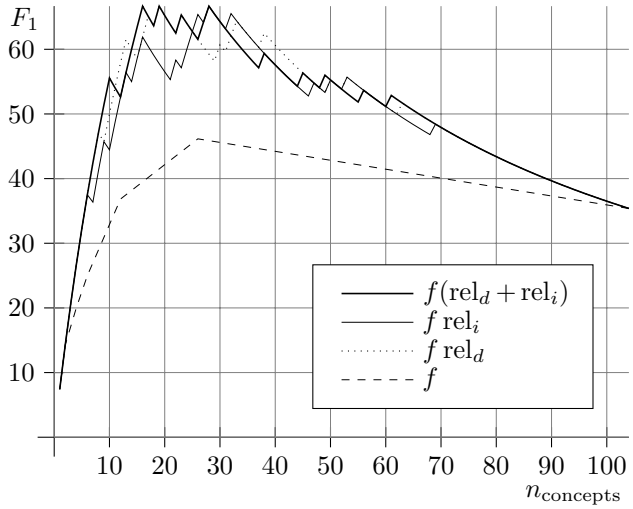

Figure 1: Selection criterion identification

The keyword selection criterion is implemented as a threshold on a relevance measure. A manually annotated document (see section 5.) has been used to identify a proper relevance measure and a proper threshold. Different relevance measures, based on different combinations of keywords frequencies f, domain relatedness rel_d and item relatedness rel_i (i.e.: the context relatedness used for disambiguation) have been tested. For each tested measure the maximum achievable F_1 value has been identified varying the threshold. In figure 1 a comparison of the best performing measures (together with the baseline f) is reported: all of them achieve the same F_1 maximum value. However $f(\text{rel}_d + \text{rel}_i)$ performs better than the other two at the surrounding threshold values. For this reason it has been selected with the thresholds 1.45 corresponding to $F_1 = 66.67\%$. In table 1 performance results on the tuning file are reported for the identified selection criterion and when no criterion is applied.

criterion	ref	hyp	match	RCL (%)	PRC (%)	F_1 (%)
none	26	104	23	**88.46**	22.12	35.38
$f(\text{rel}_d + \text{rel}_i)$	26	19	15	57.69	**78.95**	**66.67**

Table 1: Selection criterion identification results

4.3. Ontology connection

Once a keyword has been identified, disambiguated, validated and filtered, the system has to identify which concept in the ontology is the closest to the one expressed by the keyword in the text.

Although the conversion of the Wikipedia-based annotation into an ontology-based one can be performed using a map between each Wikipedia page and the most appropriate ontology concept (using methods similar to those described in (Reiter et al., 2008; Ruiz-Casado et al., 2005; Medelyan and Milne, 2008)), we preferred to take advantage of the information emerging from the documents under analysis:

a map is created between pairs (text string-Wikipedia page) and ontology concepts. This allows us to solve possible ambiguities when the same Wikipedia page refers to more than one ontology concepts, due to different granularity degrees between Wikipedia and the ontology.

4.3.1. Create the mapping table

The main assumption is that each ontology concept can be described by a synonym set (synset). The synset and ontology relations among concepts are used to create the map.
We can define the synset relatedness as the context relatedness, described in section 4.1., computed using all the terms in the synset as a context, instead of using the News items terms. The synset context can be used as disambiguation context for each term in the synset and provides disambiguation probability (in respect to the synset context) of each Wikipedia page that can be referred by that term.
In the same way, we can also define the group relatedness, for a generic set of ontology concepts, as the context relatedness evaluated using all the terms in all the synsets expressing all the concepts in the set.
The synset relatedness tells how much a Wikipedia page is related to an ontology concept and so this is the main measure used to create the map. Sometime it happens that two different concepts have the same synset relatedness value for the same Wikipedia page, as it is shown in table 2. This happens, for example, when two different ontology concepts are represented by two identical synsets. In these cases the relations among concepts are used to select the association that is more likely to be the right one. This is achieved evaluating the group relatedness for groups of concepts sharing the same relation (e.g.: being instances of the same class). The group relatedness measures how much the Wikipedia page shares the relation with the concepts in the group.
In figure 2 the map creation for two ontology concepts (turbine and water turbine) is shown. Each concept is described by a synset inside the ontology. For each term of each synset an anchor search is performed to retrieve all the related Wikipedia pages. Pages can refer to only one concept ("Water wheel", "Rotor_(electric)", ...) or to more than one concept ("Turbine", "Water turbine", ...). In the first case the only possible concept is always chosen. In the latter case synset and group relatedness, commonness and disambiguation probability are used to select the most appropriate concept. In figure 2 the page "Water turbine" can be reached by both the concepts "Turbine" and "Water turbine". However it is assigned to the concept "Water turbine" on the basis of the other measures. The term "hydraulic turbine" may disambiguate to the page "Turbine" in some contexts. If this happens, the associated ontology concept would be "Water turbine" in any case, because "hydraulic turbine" is not in the "Turbine" synset.

4.3.2. Use of the mapping table

In the previous section we've defined a map between pairs (synset term-Wikipedia page) and ontology concepts. Analogous pairs (keyword-Wikipedia page) are extracted from the text are searched in the mapping table.
When the pair is present in the table, the entry containing the pair with the highest synset relatedness for the given Wikipedia page is chosen. If more than one entry has the same synset relatedness, the other relevance measures (group relatedness, disambiguation probability and commonness) are used to select the most probable ontology concept (In table 2, for the pair "hill+Hill" the first entry is chosen). However when two ontology concepts have the same synset relatedness for a given pair it's probable that they're in fact duplicates and thus they are submitted to the ontologists for review.
In general, not all the pairs extracted from a generic text are present in the map as it may be that:

1. the keyword is not a term in the ontology, but there is at least one entry in the table that refers to the Wikipedia page of the pair. In this case the entry with the highest synset relatedness for the given Wikipedia page is chosen. If more than one entry have the same synset relatedness, the other measures are used as well.

2. the Wikipedia page in the pair cannot be found by any anchor-search performed with any ontology term. In this case the association is not possible, but the new concept may be proposed to the ontologists if the Wikipedia page seems related to the ontology domain (i.e.: the page have an high domain relatedness value).

5. Manual annotation procedure

For the evaluation of the system a set of 6 video items in the renewable energy domain (33 minutes) was manually transcribed and annotated with the most relevant concepts (keywords and associated Wikipedia pages). One of them was used to tune the system and detect the best relevance measure and threshold.
The system performance was evaluated on manual transcriptions, that can be considered generic textual documents in the renewable energy domain. Annotating video transcriptions will allow further evaluation of the performance of the system on automatically obtained transcriptions.
The manual annotation was performed independently by 3 annotators, following a common set of guidelines:

- annotate only nominal phrases (chunks) excluding quantifiers and temporal references, but keeping units of measurement. Acronyms are chunks by themselves;

- identify one or more Wikipedia pages explaining enough of the concept expressed by each chunk. Mark with a proper identifier concepts that are missing in Wikipedia;

- identify the ontology concept that is best suited to represent the meaning of each chunk;

- rank by relevance each concept on the basis of the whole text. Ranking values range from 0 to 1, meaning that the concept is not relevant at all and that it is essential, respectively. Concepts that are not central, but are part of the main story should receive a score higher than 0.5.

synset term	Wikipedia page	synset relatedness	group relatedness	disambig. prob.	common.	ontology concept	group
hill	*Hill*	0.93	**0.61**	0.83	0.91	**Hill**	Location
...	...	...	...	...	...	...	...
hill	*Hill*	0.93	0.07	0.83	0.91	**A. Hill**	Person

Table 2: Mapping table example

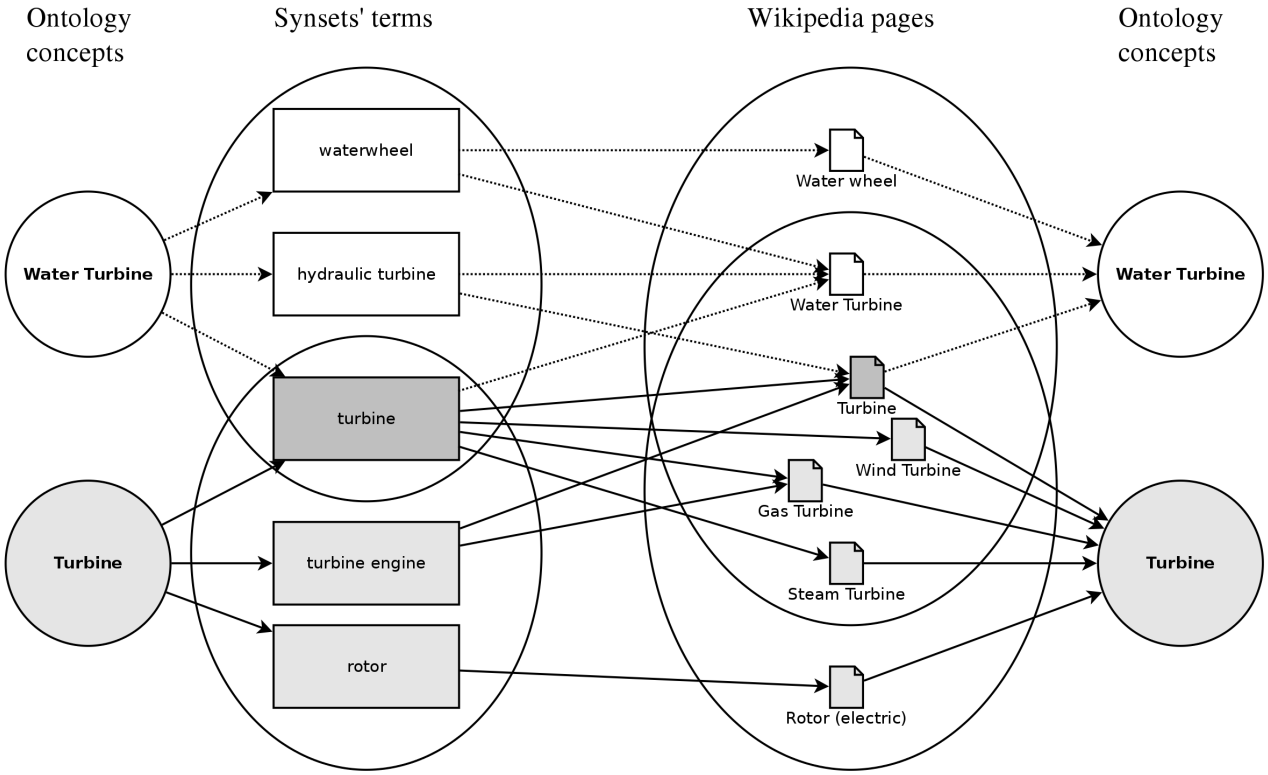

Figure 2: Anchor search for each ontology concept

The final reference annotations were created by automatically combining the 3 manual annotations. Concepts disambiguation in terms of both associated Wikipedia pages and associated ontology concepts have been discussed and agreed among all the annotators. Finally the individual relevance scores, assigned to each concept, were averaged, manually reviewed and discussed by the whole group of annotators.

6. Evaluation

The system has been evaluated on 5 manually annotated video transcriptions, accounting for 3122 words.

criterion	ref	hyp	match	RCL (%)	PRC (%)	F_1 (%)
none	84	351	64	**76.19**	18.23	29.43
$f(\text{rel}_d + \text{rel}_i) > 1.45$	84	57	41	48.81	**71.93**	**58.16**

Table 3: Wikipedia-based annotation evaluation

In table 3 results for the baseline and the implemented selection criterion are reported. The number of concepts automatically hypothesised by the system and the number of those matching the reference are also reported.

Of the 84 manually annotated concepts (with relevance above 0.7), 64 have been correctly identified (76% recall), as shown in table 3. However the system hypothesises 351 concepts, if no selection criteria is provided, resulting in a very poor precision (18%). The selection criteria implemented improves precision to 72% by selecting 57 concepts out of 351. The drawback is a loss in recall as to the initial 20 missed detections (mostly due to keyword identification and disambiguation errors) other 23 are added due to a relevance underestimation. The overall performance (F_1 score) is 58%. With respect to precision, it should be noticed that most of the "erroneously" identified concepts are actually concepts correctly identified and correctly disambiguated that don't satisfy the chosen relevance requirement (14 out of 16). Thus discrepancies between the manually assigned relevance and the system generated relatedness accounts for most precision errors (relevance overestimation) and recall errors (relevance underestimation).

With respect to this, it should be noticed that concept relevance is highly subjective and that interjudge agreement has been measured among the three annotators leading to an average Cohen's kappa coefficient of 0.45 (moderate agreement). The coefficient of agreement among the average manual relevance used for annotating concepts and the relatedness provided by the system is 0.54, falling within the confidence limits of each individual kappa coefficient. This means that disagreement between system and manually assigned relevance does not differ significantly from the assessed human disagreement.

criterion	ref	hyp	match	RCL (%)	PRC (%)	F_1 (%)
none,	46	150	37	**80.43**	24.67	37.76
$f(\mathrm{rel}_d + \mathrm{rel}_i) > 1.45$	46	42	28	60.87	**66.67**	**63.64**

Table 4: Ontology-based annotation evaluation

Among the 84 manually annotated concepts, 46 have also been manually assigned to the corresponding ontology concept (it should be noticed that the ontology used for evaluation is a domain ontology and does not cover concepts that are relevant only to the particular document context). In table 4 results for the ontology-based annotation are presented. The system automatically identifies 150 ontology concepts and automatically selects 42 of them as relevant. As 28 ontology concepts are correctly identified, the performance achieved is 64% (F_1 score).

7. Conclusions

A method has been proposed to automatically identify relevant concepts in textual documents and automatically map them to their formalization in a given domain ontology. This enables automatic annotation of texts and semantic metadata generation exploiting both Wikipedia knowledge and the Ontology knowledge. This method has already been implemented in the Papyrus (*Cultural and Historical Digital Libraries Dynamically Mined from News Archives*) prototype to provide metadata generation for the semantic search functionality and to provide content mapping to the News Ontology for the cross discipline search functionality. It analyses both textual content and speech transcripts in English and French, in two domains (renewable energy and biotechnology).

As no benchmark was available for the specific task, we manually annotated a set of documents for evaluation purposes. The performances resulted in 58% F_1 score for relevant Wikipedia pages identification and 64% for relevant ontology concepts identification.

Most errors are due to disagreement on the degree of relevance, that nevertheless fall within the inter-rater assessed disagreement. We believe that other errors actually affecting the recall measure could be further reduced by improving the candidate keyword identification process. In our future work, we plan to perform anchor search of the nominal phrases components when no match in Wikipedia is found with the whole phrase.

8. References

Olena Medelyan and David Milne. 2008. Augmenting domain-specific thesauri with knowledge from Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference (NZCSRSC 2008)*, Christchurch, NZ.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

David Milne and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*.

David Milne and Ian H. Witten. 2008b. Learning to link with Wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA. ACM.

David Milne, Olena Medelyan, and Ian H. Witten. 2006. Mining domain-specific thesauri from Wikipedia: A case study. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 442–448, Washington, DC, USA. IEEE Computer Society.

David Milne. 2009. An open-source toolkit for mining Wikipedia. Published online. Accessed 16.02.2010, 14.00h. Available from: `http://www.cs.waikato.ac.nz/~dnk2/publications/AnOpenSourceToolkitForMiningWikipedia.pdf`.

Giulio Paci, Giorgio Pedrazzi, and Roberta Turra. 2010. Wikipedia based semantic metadata annotation of audio transcripts. In *WIAMIS 2010: Eleventh International Workshop on Image Analysis for Multimedia Interactive Services*, Los Alamitos, CA, USA. IEEE Computer Society.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Published online, October. Accessed 30.10.2009, 10.30h. Available from: `http://snowball.tartarus.org/texts/introduction.html`.

Nils Reiter, Matthias Hartung, and Anette Frank. 2008. A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 381–387. College Publications.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. pages 380–386.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. Available from: `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf`.