GALATEAS D2W: A Multi-lingual Disambiguation to Wikipedia Web Service

Deirdre Lungley
CIMeC, University of Trento
Rovereto, Italy

Marco Trevisan
CELI S.R.L.
Torino, Italy

Vien Nguyen
University of Lugano
Italy

Maha Althobaiti
CSEE, University of Essex
Colchester, ESSEX, U.K.

Massimo Poesio
CIMeC, University of Trento
Rovereto, Italy

ABSTRACT

The motivation for entity extraction within a digital cultural collection is the enrichment potential of such a tool – useful in this context for such tasks as metadata generation and query log analysis. The use of Disambiguation to Wikipedia as our particular entity extraction tool is motivated by its generalisable nature and its suitability to noisy text. The particular methodolgy we use does not avail of specific natural language tools and therefore can be applied to other languages with minimal adaptation. This has allowed us to develop a multi-lingual Disambiguation to Wikipedia tool which we have deployed as a web service for the use of the community.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms,Languages, Experimentation

Keywords

Disambiguation to Wikipedia, Entity recognition

1. INTRODUCTION

Information Retrieval within the digital cultural heritage context must contend with often inately "noisy" resources: poor spelling and punctuation, obsolete word forms and abbreviated forms. This can be the case in both text-based resources and in the metadata of image-based resources. This provides an often unsurmountable challenge to traditional natural language processing techniques, e.g., traditional Named Entity Recognition (NER). However, the enrichment potential, within this context, of such a technique, e.g., metadata generation, makes it a challenge worth pursuing.

Most work on NER has focused on (a subset of) the ACE entity types [3], and in particular on **PER** (person), **LOC** (location), **ORG** (organisation). These types are very important for news but not as central in other domains: for example, in a digital art collection, types such as **WORK-OF-ART** are as important, as are finer-grained specialisations of the standard types such as **ARTIST** or **photographer**. One solution to the problem involves developing techniques for rapid domain adaptation. But such techniques are of limited use for a web service as adaptation to a domain is impossible.

The approach we adopted was to develop a service for **Disambiguation to Wikipedia**: linking text to the appropriate Wikipedia page, from which the appropriate type can then be extracted [5, 2, 6, 8]. Figure 1 shows an example of D2W. Given a text "John McCarthy, 'great man' of computer science, wins major award.", a D2W system is expected to detect the text segment "John McCarthy" and link to the correct Wikipedia page *John_Mc Carthy_(computer_scientist)*[1], instead of other *John McCarthys*, e.g., the ambassador, the senator or the linguist.

This D2W approach offers substantial coverage of entities of most domains and in multiple languages. It also allows us to contend with the noisy text environment of digital cultural collections – Wikipedia, a large encyclopedic collection from the web, allows us to extract structured knowledge which is also noisy. Our work builds on previous work in this area [4, 6], the novel aspects being our evaluation in a noisy environment and the multi-lingual adaptations.

The following section summarises previous work in this area. Section 3 details the methodology we employed to extract statistics from the structural information of Wikipedia and how these statistics are used for disambiguation. Section 4 details the steps taken to create a multi-lingual web service. We conclude with Section 5 which details an evaluation of our methodology and the future directions of this work.

ENRICH '13 Dublin, Ireland
Copyright 2013 ACM X-XXXXX-XX-X/XX/XX ...$15.00.

[1] We use the title *John_McCarthy_(computer_scientist)* to refer to the full address *http://en.wikipedia.org/wiki/John_Mc Carthy_(computer_scientist).*

2. PREVIOUS LITERATURE

A method for named entity disambiguation based on Wikipedia was presented in [2]. They first extract resources and context for each entity from the entire Wikipedia collection, then use NER in combination with other heuristics to identify named entity boundaries in articles. Finally, they employ a vector space model which includes context and categories for each entity for the disambiguation process. The approach works very well with high disambiguation accuracy. However, their use of many heuristics and NER means the method is difficult to adapt to other languages as well as to other content types such as noisy text.

A general approach for D2W is proposed by [6]. First, they process the entire Wikipedia and collect a set of incoming/outgoing links for each page. They employ a statistical method for detecting links by gathering all *n-grams* in the document and retaining those whose probability exceeds a threshold. For entity disambiguation they use machine learning with a few features, such as the commonness of a surface form, its relative relatedness in the surrounding context and the balance of these two features. Our methodology builds on this approach, adapting it for a multi-lingual environment and evaluating it on noisier text.

Local and global features are combined in an approach to the D2W task in [8]. They implement their approach using traditional *bag-of-words* and *TF-IDF* measures to calculate semantic relatedness. However, their use of many natural language specific tools, e.g., NER, chunking and part-of-speech tagging makes their method difficult to adapt to noisy text and to other languages.

Previous approaches to *D2W* differ with respect to the following aspects: 1. the corpora they address; 2. the type of the text expression they target to link; 3. the way they define and use the disambiguation context for each entity. For instance, some methods focus on linking only named entities, such as [1, 2]. The method of [2] defines the disambiguation context by using some heuristics such as entities mentioned in the first paragraph and those for which the corresponding pages refer back to the target entity. [6] utilise entities which have no ambiguous names as local context and also to compute semantic relatedness. A different method is observed in [8] where they first train a local disambiguation system and then use the prediction score of that as disambiguation context.

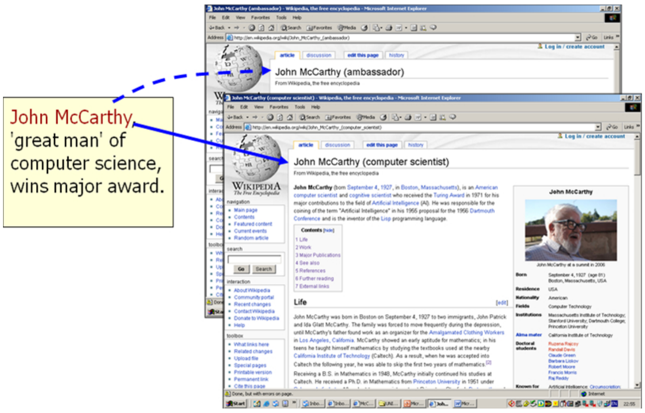
John McCarthy, 'great man' of computer science, wins major award.
John McCarthy (ambassador)
John McCarthy (computer scientist)
WIKIPEDIA

Figure 1: Disambiguation to Wikipedia

3. EXTRACTING WIKIPEDIA STRUCTURAL INFORMATION

Our approach involves two steps: first extract a number of statistical measurements from a Wikipedia dump, and then leverage these metrics in the disambiguation phase.

3.1 Parsing Wikipedia Dump

The wikipedia dump used is pages-articles.xml, published in July 2011, which contains all current edition Wikipedia articles, templates, media/file descriptions and primary meta-pages. Wikipedia parsing enables us to build necessary dictionaries and structures. All category pages, disambiguation pages, help pages, 'list_of' pages and pages referring to templates and Wikipedia itself are excluded from the parsing process. Consequently, only the most relevant pages with textual content are utilised.

In the first parse, the system scans Wikipedia articles and constructs the *redirection pairs* set (i.e., one article contains a redirect link to the actual article for that entity), and a list of all Wikipedia article titles. An example of a redirection pair is *Leonardo_da_Vinci* and *Da_Vinci*. *Leonardo_da_Vinci* is the full name of Da Vinci. Thus, the article entitled *Da_Vinci* just points to the actual article with the title *Leonardo_da_Vinci* [7].

The second parse builds a list of links, i.e., *surface form*, *target article* for each link, and the number of times one surface form is linked to the target article (incoming and outgoing links for each article are computed). We use the term 'surface form' to indicate the mention of an entity that already has a corresponding Wikipedia article and the term 'target article' to indicate the Wikipedia article that the surface form is linked to. Within the set of links, the redirected article related to the link is changed to the actual article by using the redirection pairs collected in the first parse. For example, if the title *Da_Vinci* appears in one link, we will change it to *Leonardo_da_Vinci*.

The third parse involves parsing individual Wikipedia pages to construct the: ID, title, set of categories, set of templates and set of links for each article.

Furthermore, the set of Wikipedia titles and surface forms are preprocessed byway of case-folding. So, at the end of this parsing phase, we have obtained a set of dictionaries and structures. We use the dictionaries of titles, surface forms and files (e.g, *File:Mona Lisa, by Leonardo da Vinci, from C2RMF retouched.jpg*) to match the textual content and detect entity boundaries. The set of links is used to compute statistical measures, necessary for the disambiguation phase.

3.2 Computing Statistical Measures

Links embedded in Wikipedia articles provide millions of manually defined examples to learn from, i.e., for every surface form in an article a Wikipedia editor has manually selected the correct target article that represents the meaning of the anchored text. We derive the following statistical measures from the dictionaries and structures, which we have extracted from Wikipedia.

- Keyphraseness: This is the probability that a word or a phrase will link to a Wikipedia article. Thus, to identify important words and phrases, we follow the methodology of [4] in which all word n-grams are extracted and the probability of each n-gram is computed, as follows:

$$keyphraseness(s) = \frac{count(s_{link})}{count(s)}$$

Here, s is a surface form (anchor text), $count(s_{link})$ is the number of Wikipedia articles in which s appears as a link, and *count(s)* is the number of Wikipedia articles in which s appears.

- Commonness: This is the probability of a specific Wikipedia article t to function as the target of a link with a word or a phrase s.

$$Commonness(s,t) = \frac{P(t|s)}{P(s)}$$

Here, $P(t|s)$ is the number of times s appears as a link to t, and *P(s)* is the number of times s appears as a link.

- Relatedness: This metric allows us to measure the semantic similarity of two terms [9]. We consider each term a representative Wikipedia article. For example, the term *wood* is represented by the Wikipedia page *http://en.wikipedia.org/wiki/Wood.* Pages that link to both terms suggest relatedness.

$$Relatedness(a,b) = \frac{log(max(|A|,|B|)-log(|A \cap B|))}{log(|W|)-log(min(|A|,|B|))}$$

Where a and b are the two articles of interest, A and B are the sets of all articles that link to a and b respectively. W is the entire Wikipedia.

3.3 Disambiguation Method

Our disambiguation method is made up of two steps. First, all surface forms with their potential candidates (target articles) are detected in given documents. Next, a scoring method is used to select a candidate sense, either our baseline or a machine learning method.

The identification of candidates follows the methodology employed in [4], where all word n-grams are extracted and the keyphraseness of each n-gram is computed. The keyphraseness determines the probability that each n-gram word will be a candidate to link to a Wikipedia article. Only n-grams whose keyphraseness exceed a certain threshold remain. Following preliminary experiments, we found that the best mention detection performance is accomplished with keyphraseness = 0.01.

In the next step, we employ commonness in order to identify potential candidate links. For each surface form we use the top 10 candidates (target articles) with the highest commonness. The entity disambiguation problem can be converted to a ranking problem, i.e., the link corresponding to a surface form s is defined as the one with highest score:

$$\hat{t} = \arg\max_{t_i} score(s, t_i)$$

In this formula, $score(s, t_i)$ is an appropriate scoring function.

As a baseline we use solely commonness, which is the fraction of times the title t is the target page for a surface form s. This single feature is a very reliable indicator of the correct disambiguation [8].

3.3.1 Machine Learning Method

This method uses three types of features: the commonness of each candidate link, its average relatedness in the surrounding context of the current document, and a feature which balances these two statistical measures.

The relatedness of each candidate sense is the weighted average of its relatedness to each context article as proposed by [6]:

$$score(s,t) = \frac{\sum_{c \epsilon C} relatedness(t,c)}{|C|} \times commonness(s,t)$$

where $c \epsilon C$ are the context articles of t. Only unambiguous context articles (surface forms) are considered, i.e., those that have only one related Wikipedia article. Their unambiguous nature makes them more helpful in defining the particular context.

The final feature balances commonness and relatedness, by summing the weights that were previously assigned for each unambiguous surface form, as proposed by [6].

4. BUILDING A MULTI-LINGUAL D2W WEB SERVICE

Our baseline approach has been adopted, with a few modifications, to produce seven wikifiers in seven different languages: English, Italian, French, German, Dutch, Polish and Arabic. To build each language model we acquired the language-specific Wikipedia dump and list of 'stop words'. Multi-lingual mappings of Wikipedia internal link keywords, e.g., 'category', 'help', 'file', 'disambiguation', 'template' and 'list of' were also required by the parser. UTF-8 encoding allowed for the representation of the broad range of characters required.

4.1 Web Service Technology

The D2W system has been deployed as a Web service into Linguagrid[2], a platform for the controlled distribution of NLP Web services. Software as a Service (SaaS), such as Web services, are attractive to the system integrator because they delegate the burden of the optimisation of hardware resources to the service provider. For the system provider, SaaS is also attractive since it reduces the cost of delivering and maintaining the system for multiple users and also allows him to have finer control over how the system is used. Usage statistics can also be collected and analysed.

The D2W system has been adapted and optimised to ensure it integrated effortlessly and performed efficiently as a Web service in Linguagrid. The Web service framework we used to encapsulate the D2W system into a Web service is Apache CXF. The D2W Web service exposes an API (WSDL schema) that uses data structures based on the Morphosyntactic Annotation Framework (MAF) parts, an ISO standard (ISO/DIS 24611)[3]. This WSDL schema has been developed in the context of the EUROPEANA[4] project. According to this schema, the information returned by the D2W web service consists of a set of annotations, each annotation associating the URI of a Wikipedia article to a substring of the input text. This URI can easily be used

[2] http://www.linguagrid.org

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=51934

[4] http://www.europeana.eu

within client software to retrieve linked data from ontologies and thesauri.

In order to reduce the memory requirements of GATE[5], the D2W Web service relies on a customised GATE gazetteer that reads data from a database instead of from memory. This customisation has allowed us to have a single system supporting multiple languages, at the cost of reducing the processing speed. In the context of the GALATEAS[6] project, the D2W system has been used to extract named entities from 1.5M short text queries, each 14 characters long on average (total 21M characters). The system achieved a throughput of almost 600 characters per second running on a dedicated multicore server, using a limited amount of main memory (400 MB).

5. RESULTS AND DISCUSSION

The evaluation of our D2W methodology, detailed in [7], involved 1000 queries from the Bridgeman Art Library (BAL)[7] – a noisy dataset containing spelling errors, malformed sentences, etc.. The annotators were asked to link the first five nominal mentions of each co-reference chain to Wikipedia. Table 1 details these results to highlight the potential of this methodology. The first five rows report the results for our baseline method, according to the number of disambiguation candidates generated. The last row shows the results for the machine learning method.

No. of Candidates	Recognised mention(s)	F-measure
Candidate 1	853	64.77
Candidate 2	1022	71.59
Candidate 3	1092	75.42
Candidate 4	1134	77.18
Candidate 5	1157	78.32
All features	n/a	69.32

Table 1: Results with 1000 BAL queries

A second evaluation, also detailed in [7], details the results obtained with standard datasets: ACQUAINT – a subset of this newswire text corpus which is annotated to mimic the hyperlink structure in Wikipedia, and a dataset constructed from 10,000 paragraphs from Wikipedia itself. These results, detailed in Table 2 prove the potential of the machine learnt method. Milne-Witten (2008) refers to the results reported in [6], while Ratinov-Roth refers (2011) to those in [8].

System	ACQUAINT	Wikipedia
Our M/L D2W	86.16	84.37
Milne-Witten (2008)	83.61	80.31
Ratinov-Roth (2011)	84.52	90.20

Table 2: Results with standard datasets

The novelty of our work lies in our evaluation on a noisy dataset and the conversion of this methodology to a Web service, available as a resource to the digital cultural heritage community [8]. Although the multi-lingual version of the D2W Web service has been tested and provided similar results as the original English version, a thorough evaluation remains for future work.

6. ACKNOWLEDGMENTS

This work has been supported by the GALATEAS project (http://www.galateas.eu/ – CTP-ICT PSP-2009-3-25430) funded by the European Union under the ICT PSP program.

7. REFERENCES

[1] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy, 2006.

[2] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassell, and R. Weischedel. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proc. of LREC*, 2000.

[4] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, 2008.

[5] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 233–242. ACM, 2007.

[6] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518. ACM, 2008.

[7] T.-V. T. Nguyen and M. Poesio. Entity disambiguation and linking over queries using encyclopedic knowledge. In *Proceedings of the 6th workshop on Analytics for Noisy Unstructured Text Data*, AND '12, December 2012.

[8] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384. Association for Computational Linguistics, 2011.

[9] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[5] http://gate.ac.uk/

[6] http://www.galateas.eu

[7] http://www.bridgemanart.com/

[8] http://ws.linguagrid.org/d2w