# Linking, Searching, and Visualizing Entities in Wikipedia

**Marcus Klang, Pierre Nugues**
Lund University, Department of Computer science,
Lund, Sweden
marcus.klang@cs.lth.se, pierre.nugues@cs.lth.se

**Abstract**

In this paper, we describe a new system to extract, index, search, and visualize entities in Wikipedia. To carry out the entity extraction, we designed a high-performance, multilingual, entity linker and we used a document model to store the resulting linguistic annotations. The entity linker, HEDWIG, extracts the mentions from text using a string matching engine and links them to entities with a combination of statistical rules and PageRank. The document model, Docforia (Klang and Nugues, 2017), consists of layers, where each layer is a sequence of ranges describing a specific annotation, here the entities. We evaluated HEDWIG with the TAC 2016 data and protocol (Ji and Nothman, 2016) and we reached the CEAFm scores of 70.0 on English, on 64.4 on Chinese, and 66.5 on Spanish.

We applied the entity linker to the whole collection of English and Swedish articles of Wikipedia and we used Lucene to index the layers and a search module to interactively retrieve all the concordances of an entity in Wikipedia. The user can select and visualize the concordances in the articles or paragraphs. Contrary to classic text indexing, this system does not use strings to identify the entities but unique identifiers from Wikidata. A demonstration of the entity search and visualization will be available for English at this address http://vilde.cs.lth.se:9001/en-hedwig/ and for Swedish at: http://vilde.cs.lth.se:9001/sv-hedwig/.

**Keywords:** named entity recognition, entity linker, wikipedia

## 1. Introduction

Wikipedia has become a popular NLP resource used in many projects such as text categorization (Wang et al., 2009), information extraction, question answering (Ferrucci, 2012), or translation (Smith et al., 2010). In addition to its size and diversity, Wikipedia, through its links, also enables to create a graph that associates concepts, entities, and their mentions in text. Wu and Weld (2010), for instance, used the "wikilinks", the Wikipedia hyperlinks, to collect the mentions of an entity and build sets of synonyms for an open information extraction system.

However, according to the edition rules of Wikipedia, only the first mention of an entity should be linked in an article. An automatic wikification is then necessary to associate the subsequent mentions with an entity (Mihalcea and Csomai, 2007). In addition, searching entities using names in the form of strings can be tricky as names are sometimes ambiguous and entities may have more than one name. Finding all the occurrences of an organization like the United Nations would require five or more queries as they can be mentioned not only as *the United Nations*, but also as: *UN*, *U.N.*, *United Nations Organization*, *UNO*, etc.

In this paper, we describe a novel multilingual system to process, index, search, and visualize all the mentions of an entity in Wikipedia. This system consists of an entity linker, HEDWIG, that extracts the mentions from text using a named-entity recognition engine and links them to entities with a combination of statistical rules and PageRank. We applied HEDWIG to the whole collection of English and Swedish articles of Wikipedia. We then used Lucene to index the layers and a search module to interactively retrieve all the concordances of an entity in the articles, paragraphs and metadata. The user can then select a concordance s/he wants to visualize. As opposed to the Wikipedia index, the system uses unique identifiers to index the entities and not their mentions, which enables the users to carry out more easily exhaustive searches.

## 2. Previous Work

Most named entity linkers adopt a two-step procedure, where they first identify the mentions and then link them to a unique identifier.

### 2.1. Mention Detection

The mention detection step, or spotting, has been addressed by a variety of techniques. Mihalcea and Csomai (2007) used a dictionary associating the entities with their surface forms, where the surface forms are simply n-grams. They extracted all the strings in a text that matched any of the surface forms in the dictionary to produce the set of mention candidates. As the candidates may overlap, the authors ranked them using a *keyphraseness* metric defined as the number of documents, where the mention was linked divided by the number of documents, where the mention occurred. They set the number of mentions to keep to 6% of the total number of words in the document following figures they observed in Wikipedia.

Milne and Witten (2008) also used a dictionary of surface forms as well a classifier to decide on the mentions to keep. They trained the classifier on Wikipedia mentions, either linked, the positive examples, or nonlinked, the negative ones. As features, they used the link probability (keyphraseness), relatedness, disambiguation confidence, generality, location, and spread.

Lipczak et al. (2014) used the Lucene's finite state transducers and Solr Text Tagger to detect the mentions. They collected the surface form dictionary from Wikipedia as well as Freebase and Google's wikilinks. The tagging step results in an overdetection that is pruned using lexical filters. The final selection of mentions is carried out in the linking step.

Cucerzan (2014) used a dictionary of surface forms collected from Wikipedia, anchor text, page titles, redirection pages, etc, and a set of rules to identify the mentions in the text. As in Lipczak et al. (2014), the overgeneration is

solved at the linking stage.
Piccinno and Ferragina (2014) used a dictionary of surface forms similar to Cucerzan (2014) to spot the mentions. They also used a pruner to discard unlikely annotations based on a classifier and a coherence metric with the set of neighboring entities. This final selection is done at linking time.
Sil et al. (2015) used classifiers based on neural nets and conditional random fields trained on three languages.
Some annotators also used an external named entity recognition module to carry out this mention detection as AIDA (Hoffart et al., 2011) and Tan et al. (2015) that used Stanford NER (Finkel et al., 2005).

### 2.2. Entity Linking

Bagga and Baldwin (1998) is one of the earliest works that introduced the notion of linkage to unique things through the task of cross-document coreference. The main difference with entity linking is that predefined lists of entities do not exist but have to be found. Bagga and Baldwin (1998) created summary vectors and tried to cluster them to form linkages. These summary vectors were created from noun phrases contained within coreference chains in documents. Using cosine similarity with a predefined threshold, they were able to cluster coreferences that crossed the document boundaries.
Bunescu and Pasca (2006) first explored entity linking using Wikipedia as knowledge base. They used hyperlinks, redirects, disambiguation pages, and the category hierarchy, which would be used by almost every major paper since. Using context article similarity based on 55-word window vector space model (VSM) cosine similarity and a taxonomy kernel, they trained SVM models to recast the disambiguation problem as a classification. They reported accuracies between 55.4% and 84.8% depending on which model they used.
Cucerzan (2007a) introduced clearly defined end-to-end pipelines – starting with text and ending with linked entities – as well as a notion of collective agreement in the disambiguation component. Using a *document vector* comprised of surface form context, entity context, and categories, he could maximize an agreement between the proposed entity candidates. Using the top two stories from 10 MSNBC news categories in January 2, 2007, he reported an accuracy of 91.4% versus 88.3% from 350 random Wikipedia pages.
Milne and Witten (2008) introduced important concepts such as relatedness and commonness which still defines a strong baseline used by many following papers in one form or another.
Hoffart et al. (2011) used an ensemble system to compute a linear combination of entity probabilities, context similarities, and entity coherences, where the *popularity prior* corresponds to the number of in-links to a Wikipedia entity; the *context similarity* compares the context of the input by computing a similarity between all the tokens in the input against a key phrase defined for entities they extracted from YAGO. A key phrase is a phrase that is derived from link texts, category names, citation titles, and other references; finally, *coherence* provides a way of comparing different entity candidates in a text in order to measure how compatible they are.
Lipczak et al. (2014) built a set of all the entity candidates for all the mentions in a document. They started from an entity core corresponding to the default senses. Using this core, they built a topic centroid from Wikipedia categories and discard entities from the core that are outside the topic. They finally refined the core and rank the remaining entities using a cosine similarity.
Eckhardt et al. (2014) built a graph of entity-mention pairs, where they weighted the edges with $P(E|M)$ probabilities. They applied a variant to PageRank to rank the entities.
Sil et al. (2015) described a trilingual system that uses a classifier with features such as the number of mention–entity matches in Wikipedia, acronym match, pointwise mutual information between entities and categories, etc.
Tan et al. (2015) used a graph of entity-mention and entity-entity edges. The edges are weighted by a function of the context similarity between a mention and an entity description in Freebase and functions of relatedness and context similarities. The entity ranking is eventually determined by a random walk in the graph.
Cucerzan (2007b) and Han and Zhao (2009) described other algorithms for NERL. In contrast to most of these previous works, multilingual support is at the core of HEDWIG.

## 3. Extraction of the Wikipedia Structure

Before we apply the linker to Wikipedia, we convert the HTML pages into a multilayer document model; see Sect. 5. This preprocessing step parses the HTML documents into DOM trees and extracts the original page structure, text styles, links, lists, and tables. We then resolve all the Wikipedia links to unique Wikidata identifiers, where Wikidata is an entity database, which assigns unique identifiers across all the language editions of Wikipedia.
The United Nations, for instance, has the unique id: Q1065, which enables to retrieve the article pages in English, French, Swedish, or Russian. Figure 1 shows examples of these ids in the *United Nations* article from the English Wikipedia, where we have replaced the manually set Wikipedia anchors (the wikilinks) with their Wikidata numbers: Q245065 for *intergovernmental organization* and Q60 for *New York City*. Figure 2 shows the first paragraph of the corresponding article from the Swedish Wikipedia, *Förenta nationerna* 'United Nations', where *mellanstatlig organisation*, the Swedish word for intergovernmental organization, has also the Q245065 number.

## 4. Entity Linking

Once we have collected and structured the text, we apply the entity linking module to find all the mentions of an entity in text and link these mentions to a unique identifier.

### 4.1. Set of Entities

We used the wikilinks to build a repository of (mention, entity) pairs and Wikidata as the nomenclature for the unique entity identifiers. We collected all the wikilinks in the Wikipedia articles, where each link consists of a label and the name of the destination page:

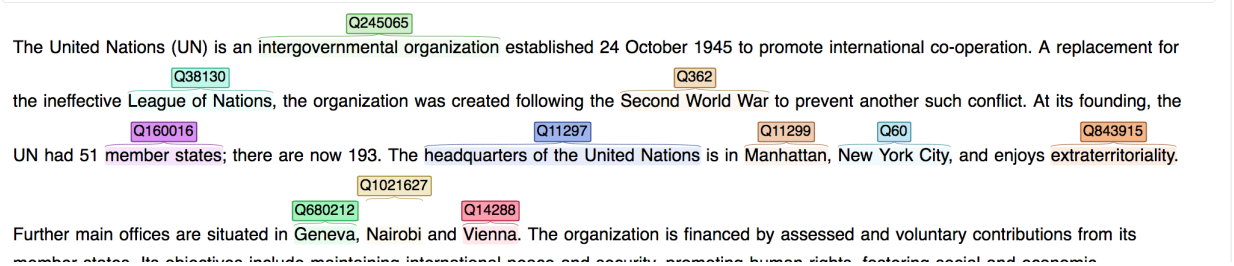



Figure 1: Visualization of anchors with Wikidata Q-numbers. The first lines of the *United Nations* article in the English Wikipedia

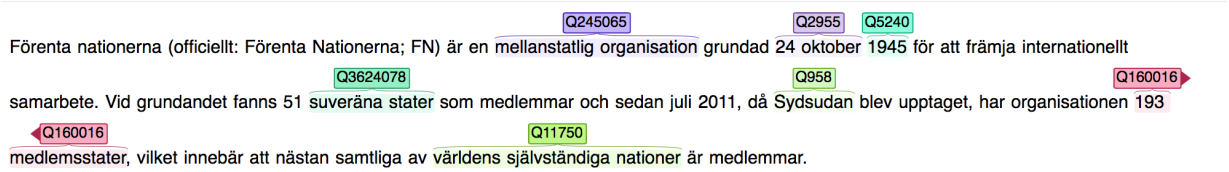



Figure 2: Visualization of anchors with Wikidata Q-numbers. First paragraph of the *Förenta nationerna* 'United Nations' article in the Swedish Wikipedia

`[[destination|label]]`. We parsed these links into (mention, entity page) pairs and we translated the entity pages into Wikidata Q-numbers.

We annotated each mention-entity pair with a set of properties: its frequency, its frequency relative to the mention, $P(E|M)$, if the mention is in a dictionary, if the mention consists of stop words. We then pruned the knowledge base from unique mentions for entities with a high frequency, mentions that are only stop words, etc.

During the mention gathering, we also derived statistics for a given language. Before we computed these statistics, we applied a procedure that we called *autolinking*. In an article, the Wikipedia guidelines advise to link only one instance of an entity mention[1]: Normally the first one in the text. With the autolinking procedure, we link all the remaining mentions provided that we have sequences of exactly matching tokens.

The statistics we collect are:

- The frequency of the mention string over the whole Wikipedia collection (restricted to one language);
- The frequency of the pair (mention, entity) that we derive from the links without autolinking (only manually linked mentions);
- The count of (entity1, entity2) pairs in a window corresponding to a paragraph and limited to 20 linked mentions. This is carried out after autolinking;
- Capitalization statistics for all the tokens: We extract token counts for all tokens with a frequency greater than 100 and we break them down by case properties: uppercased, lowercased, titlecased, and camelcase;

[1] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#Overlinking_and_underlinking`

### 4.2. Mention Recognition

To detect the mentions in an unannotated text, we use a two-step procedure: We first generate the mention candidates using a finite-state transducer; this results in a very large overgeneration. We then apply a mention segmenter that identifies the mentions to keep for the linking phase.

Following Lipczak et al. (2014) and Södergren and Nugues (2017), we used an automaton to spot the mentions. This automaton uses Lucene's finite-state transducers and is efficient in terms of memory usage and execution time. Depending on the language and the availability of manually-annotated data, we can complement this candidate generation with two named-entity recognition systems trained on the annotated data: The first one being based on an extension of the fixed-size ordinally forgetting encoding (FOFE) technique (Xu et al., 2017; Zhang et al., 2015) and the second one being CoreNLP (Manning et al., 2014).

The overgeneration of mention candidates impairs the quality of the downstream linker. To discard the very unlikely ones, we introduced rules based on the frequency of the manual links applied to mention $M$ and its link probability $lp$. We denote $M_{linked}$ a mention with a manual hyperlink; this would correspond to the wiki markup: `[[link|mention]]`, and $M_{autolinked}$, an autolinked mention. We define:

$$\begin{aligned} lp(M) &= P(M_{autolinked}|M) \\ &= \frac{\#M_{autolinked}}{\#M_{autolinked} + \#M_{unlinked}}, \end{aligned}$$

where $\#M_{autolinked}$ is the number of times a mention is linked in the Wikipedia collection and $\#M_{unlinked}$, its frequency when unlinked.

The rules are:

1. Remove the mentions $M$ where $lp(M) < D_{lp}$, for instance with $D_{lp} = 0.01$;

2. Keep the mentions where $lp(M) > K_{lp}$, and $\#M_{linked} > K_f$, with, for instance, $K_{lp} = 0.15$ and $K_f = 25$. All these mentions are candidates for the linking step;

3. Set the rest in a dubious set.

### 4.3. The Linking Step

We applied the JUNG implementation of PageRank (Brin and Page, 1998; O'Madadhain et al., 2003) to the tagged mentions. Following Eckhardt et al. (2014) and Södergren and Nugues (2017), we created a node for every mention-entity pair that is detected in the text and we ran PageRank on this graph; we used the JUNG default settings.

We analyzed the internal links of Wikipedia to determine the entities that appear in the same context. Two entities are linked if the article of Entity $A$ links to the article of Entity $B$ or there exist at least one link to the article of Entity $A$ and another one to the article of Entity $B$ occurring in the same paragraph.

We then re-ranked the PageRank candidates using a feed forward neural network consisting of three layers with RELU activations, a crossentropy loss, and a sigmoid output. We trained the model on the output of the PageRank disambiguator applied to the TAC 2015 dataset. The features we used consist of the mention tokens, candidate title tokens, both as word embeddings on 256 dimensions, the Jaccard distance between the mention and candidate title, the commonness and pagerank weights.

We evaluated the system with the same method as used in the TAC 2016 competition (Ji and Nothman, 2016) and we reached the `CEAFm` scores of 70.0 on English, on 64.4 on Chinese, and 66.5 on Spanish. We applied our linker to Swedish without any language adaptation.

We deployed the entity linker on our cluster and we used HDFS to spread the Wikipedia dump across the nodes as well as to save the final result.

## 5. The Document Model

We represented Wikipedia and the entity annotations using the Docforia document model[2] (Klang and Nugues, 2016b; Klang and Nugues, 2016a; Klang and Nugues, 2017). Docforia is designed it so that we can store the original markup, as well as any subsequent linguistic annotation. It consists of multiple layers, where each layer is dedicated to a specific type of annotation.

The annotations are encoded in the form of graph nodes, where a node represents a piece of data: a token, a sentence, a named entity, etc., delimited by ranges. These nodes are possibly connected by edges as in dependency graphs. The data structure used is similar to a property graph.

## 6. Indexing

We created an indexing tool, Panforia, to retrieve the entities from the annotated documents. As input, Panforia uses the output of the entity annotation in the form of Parquet files. Panforia is based on the Lucene search and indexing library. Each Docforia record is converted into a Lucene document by mapping record properties and documents to Lucene fields. In addition, a binary copy of the Docforia record is embedded with the indexed fields, which provides the ranges and relationships between nodes needed for the visualization.

Building directly on the Lucene library, instead of existing packages such as Solr or ElasticSearch, allowed us to optimize the insertion into an index. One key advantage of the Panforia indexer is that it can read the output from the Wikipedia pipeline, Parquet files, without a conversion step.

## 7. Visualization

The front-end of Panforia is a web server that embeds the Docforia library, Lucene, and a client-side web application. To search an entity, we enter a Wikidata Q-number, for instance, `urn:wikidata:Q168756`, corresponding to the entity identifier, here the University of California, Berkeley. Figure 3 shows the results of this search, where in each row, the entity is listed by its mention together with its left and right contexts. The document that contains the source of the concordance is listed in the leftmost column and the offset from the beginning in the last column.

In the figure, we can see that the entity has many possible mentions: *University of California, Berkeley*, *Berkeley*, *UC Berkeley*, etc. All these mentions and concordances are automatically retrieved through the entity index. We can visualize the document by clicking on a link in the left column.

For each document, the interactive visualization tool also enables the user to examine the annotated layers resulting from the HTML parsing (Sect. 3.). These layers include the manually set anchors, the automatically detected entities, and text enrichment. These layers are selectable from the dropdown menu to the right. Figure 4 shows an example with the automatically linked entities, the text in bold (strong) and in italics.

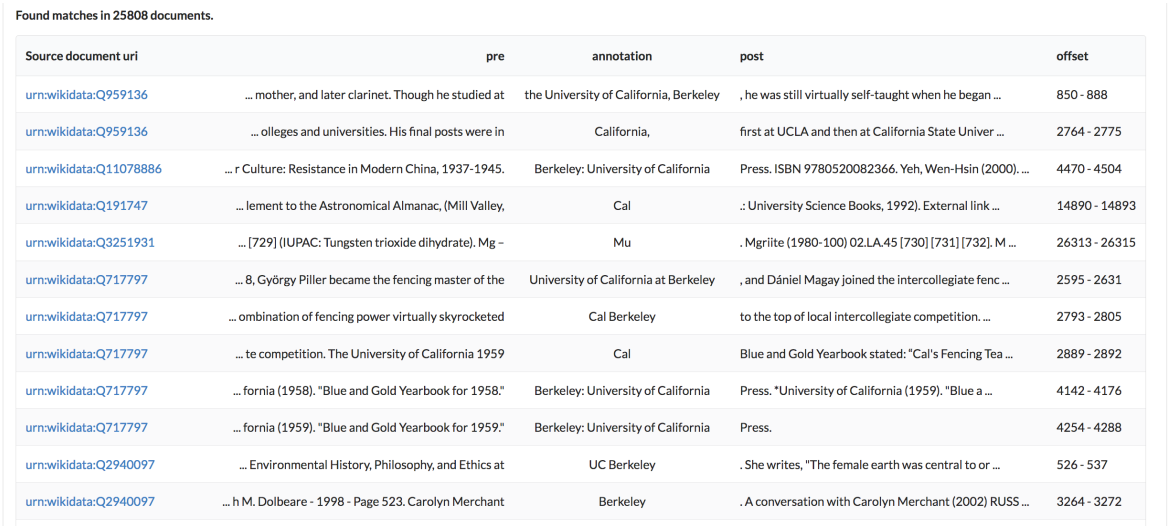
Figure 5 shows an example of results we obtained in the Swedish Wikipedia when we searched the entity Göran Persson, the former Swedish Prime Minister, using his Q-number: Q53747. This mention, *Göran Persson*, is ambiguous and Wikipedia lists as many as four different entities with this name: The former Swedish Prime Minister, a progressive musician (Q6042900), a Swedish social democratic politician, former member of the Riksdag (Q5626648), and a Swedish statesman from the 16th century (Q2625684). The latter is also spelled Jöran Person.

Searching the mention *Göran Persson* would return articles or concordances with any of these entities, while searching the entity through its Q-number only returns the intended person, either with her/his name or with other mentions such as *Persson* or *Göran*. The results are given in the forms of concordances with left and right contexts (Fig. 5).

## 8. Conclusion and Future Work

We have described a system to extract, index, search, and visualize entities on the English and Swedish Wikipedia. Given a Wikidata Q-number, a user can interactively retrieve all the concordances of an entity in the articles, para-

[2] `https://github.com/marcusklang/docforia/`



Found matches in 25808 documents.

| Source document uri | pre | annotation | post | offset |
|---|---|---|---|---|
| urn:wikidata:Q959136 | ... mother, and later clarinet. Though he studied at | the University of California, Berkeley | , he was still virtually self-taught when he began ... | 850 - 888 |
| urn:wikidata:Q959136 | ... olleges and universities. His final posts were in | California, | first at UCLA and then at California State Univer ... | 2764 - 2775 |
| urn:wikidata:Q11078886 | ... r Culture: Resistance in Modern China, 1937-1945. | Berkeley: University of California | Press. ISBN 9780520082366. Yeh, Wen-Hsin (2000). ... | 4470 - 4504 |
| urn:wikidata:Q191747 | ... lement to the Astronomical Almanac, (Mill Valley, | Cal | .: University Science Books, 1992). External link ... | 14890 - 14893 |
| urn:wikidata:Q3251931 | ... [729] (IUPAC: Tungsten trioxide dihydrate). Mg – | Mu | . Mgriite (1980-100) 02.LA.45 [730] [731] [732]. M ... | 26313 - 26315 |
| urn:wikidata:Q717797 | ... 8, György Piller became the fencing master of the | University of California at Berkeley | , and Dániel Magay joined the intercollegiate fenc ... | 2595 - 2631 |
| urn:wikidata:Q717797 | ... ombination of fencing power virtually skyrocketed | Cal Berkeley | to the top of local intercollegiate competition. ... | 2793 - 2805 |
| urn:wikidata:Q717797 | ... te competition. The University of California 1959 | Cal | Blue and Gold Yearbook stated: "Cal's Fencing Tea ... | 2889 - 2892 |
| urn:wikidata:Q717797 | ... fornia (1958). "Blue and Gold Yearbook for 1958." | Berkeley: University of California | Press. *University of California (1959). "Blue a ... | 4142 - 4176 |
| urn:wikidata:Q717797 | ... fornia (1959). "Blue and Gold Yearbook for 1959." | Berkeley: University of California | Press. | 4254 - 4288 |
| urn:wikidata:Q2940097 | ... Environmental History, Philosophy, and Ethics at | UC Berkeley | . She writes, "The female earth was central to or ... | 526 - 537 |
| urn:wikidata:Q2940097 | ... h M. Dolbeare - 1998 - Page 523. Carolyn Merchant | Berkeley | . A conversation with Carolyn Merchant (2002) RUSS ... | 3264 - 3272 |

Figure 3: Searching an entity in the Wikipedia pages, where Q168756 is the Wikidata identifier of the University of California, Berkeley. The entity concordances, where each concordance is listed with its source, mention in the text, left and right contexts, and position in the text

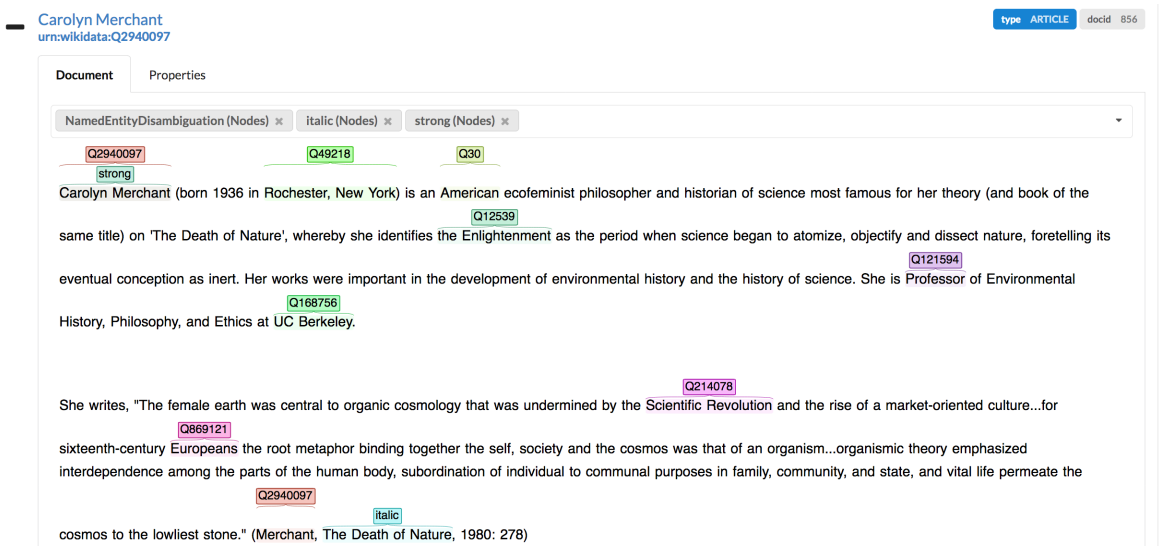
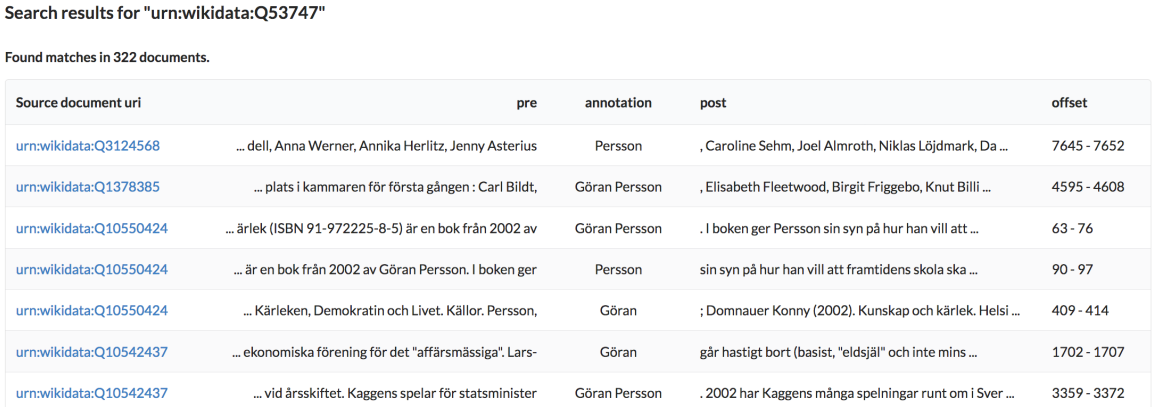



Figure 4: Visualization of annotated layers: The automatically linked entities, text in bold and in italics



**Search results for "urn:wikidata:Q53747"**

Found matches in 322 documents.

| Source document uri | pre | annotation | post | offset |
|---|---|---|---|---|
| urn:wikidata:Q3124568 | ... dell, Anna Werner, Annika Herlitz, Jenny Asterius | Persson | , Caroline Sehm, Joel Almroth, Niklas Löjdmark, Da ... | 7645 - 7652 |
| urn:wikidata:Q1378385 | ... plats i kammaren för första gången : Carl Bildt, | Göran Persson | , Elisabeth Fleetwood, Birgit Friggebo, Knut Billi ... | 4595 - 4608 |
| urn:wikidata:Q10550424 | ... ärlek (ISBN 91-972225-8-5) är en bok från 2002 av | Göran Persson | . I boken ger Persson sin syn på hur han vill att ... | 63 - 76 |
| urn:wikidata:Q10550424 | ... är en bok från 2002 av Göran Persson. I boken ger | Persson | sin syn på hur han vill att framtidens skola ska ... | 90 - 97 |
| urn:wikidata:Q10550424 | ... Kärleken, Demokratin och Livet. Källor. Persson, | Göran | ; Domnauer Konny (2002). Kunskap och kärlek. Helsi ... | 409 - 414 |
| urn:wikidata:Q10542437 | ... ekonomiska förening för det "affärsmässiga". Lars- | Göran | går hastigt bort (basist, "eldsjäl" och inte mins ... | 1702 - 1707 |
| urn:wikidata:Q10542437 | ... vid årsskiftet. Kaggens spelar för statsminister | Göran Persson | . 2002 har Kaggens många spelningar runt om i Sver ... | 3359 - 3372 |

Figure 5: Concordances of the entity *Göran Person*, Q53747. The results are given in the form of concordances with a left and right contexts.

graphs, and metadata. The user can then select a concordance and the annotations s/he wants to visualize.

This system could be improved in many ways. The entity linker makes no assumption on the language and could easily be applied to other Wikipedias. We plan to extend this demonstration to four other languages: French, German, Spanish, and Russian and for one entity, show the concordances in the six languages.

Finally, we plan to introduce a coreference resolution for the languages where a coreference-annotated corpus exists or where a solver is available.

The demonstrations will be available at: `http://vilde.cs.lth.se:9001/en-hedwig/` for English and `http://vilde.cs.lth.se:9001/sv-hedwig/` for Swedish.

## 9. Acknowledgements

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

## 10. Bibliographical References

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April.

Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *European Chapter of the Association for Computational Linguistics*, volume 6, pages 9–16.

Cucerzan, S. (2007a). Large-scale named entity disambiguation based on wikipedia data. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, volume 7, pages 708–716.

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

Cucerzan, S. (2014). Name entities made obvious: The participation in the ERD 2014 evaluation. In *Proceedings of Entity Recognition and Disambiguation, ERD'14*, Gold Coast.

Eckhardt, A., Hreško, J., Procházka, J., and Smrž, O. (2014). Entity linking based on the co-occurrence graph and entity probability. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 37–44.

Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor.

Han, X. and Zhao, J. (2009). Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 215 – 224.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh.

Ji, H. and Nothman, J. (2016). Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End KBP. In *Proceedings of the Ninth Text Analysis Conference (TAC 2016)*, Gaithersburg, Maryland. National Institute of Standards and Technology.

Klang, M. and Nugues, P. (2016a). Docforia: A multilayer document model. In *Proceedings of SLTC 2016*, Umeå, November.

Klang, M. and Nugues, P. (2016b). WIKIPARQ: A tabulated Wikipedia resource using the Parquet format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4141–4148, Portorož, Slovenia, may.

Klang, M. and Nugues, P. (2017). Docforia: A multilayer document model. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, page 226–230, Gothenburg, May.

Lipczak, M., Koushkestani, A., and Milios, E. (2014). Tulip: Lightweight entity recognition and disambiguation using wikipedia-based topic centroids. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 31–36.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on CIKM*, CIKM '07, pages 233–242, Lisbon, Portugal.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518.

O'Madadhain, J., Fisher, D., White, S., and Boey, Y.-B. (2003). The JUNG (Java Universal Network/Graph) framework. Technical Report UCI-ICS 03-17, School of Information and Computer Science, University of California, Irvine.

Piccinno, F. and Ferragina, P. (2014). From tagme to wat: A new entity annotator. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 55–62, New York, NY, USA. ACM.

Sil, A., Dinu, G., and Florian, R. (2015). The IBM system for trilingual entity discovery and linking at TAC 2015. In *Proceedings of the Eighth Text Analysis Conference (TAC2015)*.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 403–411.

Södergren, A. and Nugues, P. (2017). A multilingual entity linker using PageRank and semantic graphs. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, page 87–95, Gothenburg, May.

Tan, Y., Zheng, D., Li, M., and Wang, X. (2015). BUPT-Team participation at TAC 2015 knowledge base population. In *Proceedings of the Eighth Text Analysis Conference (TAC2015)*.

Wang, P., Hu, J., Zeng, H.-J., and Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3):265–281.

Wu, F. and Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the ACL*, page 118–127, Uppsala, Sweden.

Xu, M., Jiang, H., and Watcharawittayakul, S. (2017). A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1247, Vancouver, Canada, July. Association for Computational Linguistics.

Zhang, S., Jiang, H., Xu, M., Hou, J., and Dai, L. (2015). The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 495–500, Beijing, China, July. Association for Computational Linguistics.