Large-Scale Named Entity Disambiguation Based on Wikipedia Data

Silviu Cucerzan
Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
silviu@microsoft.com

Abstract

This paper presents a large-scale system for the recognition and semantic disambiguation of named entities based on information extracted from a large encyclopedic collection and Web search results. It describes in detail the disambiguation paradigm employed and the information extraction process from Wikipedia. Through a process of maximizing the agreement between the contextual information extracted from Wikipedia and the context of a document, as well as the agreement among the category tags associated with the candidate entities, the implemented system shows high disambiguation accuracy on both news stories and Wikipedia articles.

1 Introduction and Related Work

The ability to identify the *named entities* (such as people and locations) has been established as an important task in several areas, including topic detection and tracking, machine translation, and information retrieval. Its goal is the identification of mentions of entities in text (also referred to as *surface forms* henceforth), and their labeling with one of several entity type labels. Note that an entity (such as George W. Bush, the current president of the U.S.) can be referred to by multiple surface forms (e.g., "George Bush" and "Bush") and a surface form (e.g., "Bush") can refer to multiple entities (e.g., two U.S. presidents, the football player Reggie Bush, and the rock band called Bush).

When it was introduced, in the 6th Message Understanding Conference (Grishman and Sundheim, 1996), the named entity recognition task comprised three entity identification and labeling subtasks: ENAMEX (proper names and acronyms designating persons, locations, and organizations), TIMEX (absolute temporal terms) and NUMEX (numeric expressions, monetary expressions, and percentages). Since 1995, other similar named entity recognition tasks have been defined, among which CoNLL (e.g., Tjong Kim Sang and De Meulder, 2003) and ACE (Doddington et al., 2004). In addition to structural disambiguation (e.g., does "the Alliance for Democracy in Mali" mention one, two, or three entities?) and entity labeling (e.g., does "Washington went ahead" mention a person, a place, or an organization?), MUC and ACE also included a within document coreference task, of grouping all the mentions of an entity in a document together (Hirschman and Chinchor, 1997).

When breaking the document boundary and scaling entity tracking to a large document collection or the Web, resolving semantic ambiguity becomes of central importance, as many surface forms turn out to be ambiguous. For example, the surface form "Texas" is used to refer to more than twenty different named entities in Wikipedia. In the context "former Texas quarterback James Street", Texas refers to the University of Texas at Austin; in the context "in 2000, Texas released a greatest hits album", Texas refers to the British pop band; in the context "Texas borders Oklahoma on the north", it refers to the U.S. state; while in the context "the characters in Texas include both real and fictional explorers", the same surface form refers to the novel written by James A. Michener.

Bagga and Baldwin (1998) tackled the problem of cross-document coreference by comparing, for any pair of entities in two documents, the word vectors built from all the sentences containing mentions of the targeted entities. Ravin and Kazi (1999) further refined the method of solving coreference through measuring context similarity and integrated it into *Nominator* (Wacholder et al., 1997), which was one of the first successful systems for named entity recognition and co-reference resolution. However, both studies targeted the clustering of all mentions of an entity across a given document collection rather than the mapping of these mentions to a given reference list of entities.

A body of work that did employ reference entity lists targeted the resolution of geographic names in

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716, Prague, June 2007.

text. Woodruff and Plaunt (1994) used a list of 80k geographic entities and achieved a disambiguation precision of 75%. Kanada (1999) employed a list of 96k entities and reported 96% precision for geographic name disambiguation in Japanese text. Smith and Crane (2002) used the Cruchley's and the Getty thesauri, in conjunction with heuristics inspired from the Nominator work, and obtained between 74% and 93% precision at recall levels of 89-99% on five different history text corpora. Overell and Rüger (2006) also employed the Getty thesaurus as reference and used Wikipedia to develop a co-occurrence model and to test their system.

In many respects, the problem of resolving ambiguous surface forms based on a reference list of entities is similar to the lexical sample task in word sense disambiguation (WSD). This task, which has supported large-scale evaluations – SENSEVAL 1-3 (Kilgarriff and Rosenzweig, 2000; Edmonds and Cotton, 2001; Mihalcea et al., 2004) – aims to assign dictionary meanings to all the instances of a predetermined set of polysemous words in a corpus (for example, choose whether the word "church" refers to a building or an institution in a given context). However, these evaluations did not include proper noun disambiguation and omitted named entity meanings from the targeted semantic labels and the development and test contexts (e.g., "Church and Gale showed that the frequency [..]").

The problem of resolving ambiguous names also arises naturally in Web search. For queries such as "Jim Clark" or "Michael Jordan", search engines return blended sets of results referring to many different people. Mann and Yarowsky (2003) addressed the task of clustering the Web search results for a set of ambiguous personal names by employing a rich feature space of biographic facts obtained via bootstrapped extraction patterns. They reported 88% precision and 73% recall in a three-way classification (most common, secondary, and other uses).

Raghavan et al. (2004) explored the use of entity language models for tasks such as clustering entities by profession and classifying politicians as liberal or conservative. To build the models, they recognized the named entities in the TREC-8 corpus and computed the probability distributions over words occurring within a certain distance of any instance labeled as Person of the canonical surface form of 162 famous people.

Our aim has been to build a named entity recognition and disambiguation system that employs a comprehensive list of entities and a vast amount of world knowledge. Thus, we turned our attention to the Wikipedia collection, the largest organized knowledge repository on the Web (Remy, 2002).

Wikipedia was successfully employed previously by Strube and Ponzetto (2006) and Gabrilovich and Markovitch (2007) to devise methods for computing semantic relatedness of documents, WikiRelate! and Explicit Semantic Analysis (ESA), respectively. For any pair of words, WikiRelate! attempts to find a pair of articles with titles that contain those words and then computes their relatedness from the word-based similarity of the articles and the distance between the articles' categories in the Wikipedia category tree. ESA works by first building an inverted index from words to all Wikipedia articles that contain them. Then, it estimates a relatedness score for any two documents by using the inverted index to build a vector over Wikipedia articles for each document and by computing the cosine similarity between the two vectors.

The most similar work to date was published by Bunescu and Paşca (2006). They employed several of the disambiguation resources discussed in this paper (Wikipedia entity pages, redirection pages, categories, and hyperlinks) and built a context-article cosine similarity model and an SVM based on a taxonomy kernel. They evaluated their models for person name disambiguation over 110, 540, and 2,847 categories, reporting accuracies between 55.4% and 84.8% on (55-word context, entity) pairs extracted from Wikipedia, depending on the model and the development/test data employed.

The system discussed in this paper performs both named entity identification and disambiguation. The entity identification and in-document coreference components resemble the Nominator system (Wacholder et al., 1997). However, while Nominator made heavy use of heuristics and lexical clues to solve the structural ambiguity of entity mentions, we employ statistics extracted from Wikipedia and Web search results. The disambiguation component, which constitutes the main focus of the paper, employs a vast amount of contextual and category information automatically extracted from Wikipedia over a space of 1.4 million distinct entities/concepts, making extensive use of the highly interlinked structure of this collection. We augment the Wikipedia category information with information automatically extracted from Wikipedia list pages and use it in conjunction with the context information in a vectorial model that employs a novel disambiguation method.

2 The Disambiguation Paradigm

We present in this section an overview of the proposed disambiguation model and the world knowledge data employed in the instantiation of the model discussed in this paper. The formal model is discussed in detailed in Section 5.

The world knowledge used includes the known entities (most articles in Wikipedia are associated to an entity/concept), their entity class when available (Person, Location, Organization, and Miscellaneous), their known surface forms (terms that are used to mention the entities in text), contextual evidence (words or other entities that describe or co-occur with an entity), and category tags (which describe topics to which an entity belongs to).

For example, Figure 1 shows nine of the over 70 different entities that are referred to as "Columbia" in Wikipedia and some of the category and contextual information associated with one of these entities, the Space Shuttle Columbia.

The disambiguation process uses the data associated with the known surface forms identified in a document and all their possible entity disambiguations to *maximize* the agreement between the context data stored for the candidate entities and the contextual information in the document, and also, the agreement among the category tags of the candidate entities. For example, a document that contains the surface forms "Columbia" and "Discovery" is likely to refer to the Space Shuttle Columbia and the Space Shuttle Discovery because these candidate entities share the category tags *LIST_astronomical_topics*, *CAT_Manned_spacecraft*, *CAT_Space_Shuttles* (the extraction of such tags is presented in Section 3.2), while other entity disambiguations, such as Columbia Pictures and Space Shuttle Discovery, do not share any common category tags. The agreement maximization process is discussed in depth in Section 5.

This process is based on the assumption that typically, all instances of a surface form in a document have the same meaning. Nonetheless, there are a non-negligible number of cases in which the *one sense per discourse* assumption (Gale et al., 1992) does not hold. To address this problem, we employ an iterative approach, of shrinking the context size used to disambiguate surface forms for which there is no *dominating* entity disambiguation at document level, performing the disambiguation at the paragraph level and then at the sentence level if necessary.

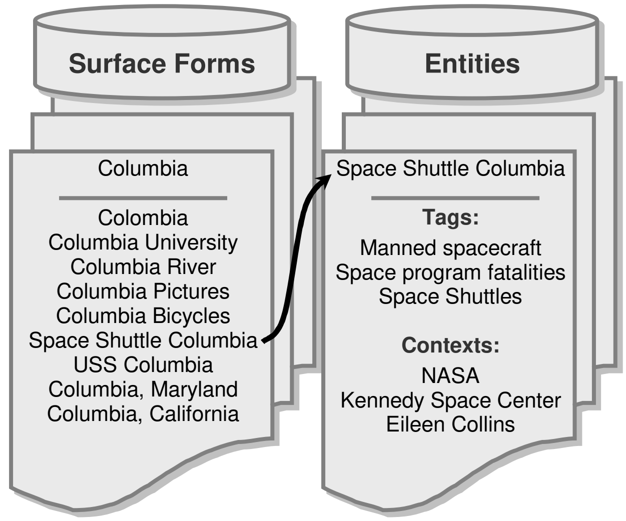

Figure 1. The model of storing the information extracted from Wikipedia into two databases.

3 Information Extraction from Wikipedia

We discuss now the extraction of entities and the three main types of disambiguation clues (entity surface forms, category tags, and contexts) used by the implemented system. While this information extraction was performed on the English version of the Wikipedia collection, versions in other languages or other collections, such as Encarta or WebMD, could be targeted in a similar manner.

When processing the Wikipedia collection, we distinguish among four types of articles: *entity pages*, *redirecting pages*, *disambiguation pages*, and *list pages*. The characteristics of these articles and the processing applied to each type to extract the three sets of clues employed by the disambiguation model are discussed in the next three subsections.

3.1 Surface Form to Entity Mappings

There are four sources that we use to extract entity surface forms: the titles of entity pages, the titles of redirecting pages, the disambiguation pages, and the references to entity pages in other Wikipedia articles. An *entity page* is an article that contains information focused on one single entity, such as a person, a place, or a work of art. For example, Wikipedia contains a page titled "Texas (TV series)", which offers information about the soap opera that aired on NBC from 1980 until 1982. A *redirecting page* typically contains only a reference to an entity page. For example, the article titled "Another World in Texas" contains a redirection to the article titled "Texas (TV series)". From these two articles, we extract the entity Texas (TV series) and its surface forms *Texas (TV series)*, *Texas* and *Another World in Texas*. As shown in this example, we store not only the exact article titles but also the corresponding forms from which we eliminate appositives (either within parentheses or following a comma).

We also extract surface form to entity mappings from Wikipedia disambiguation pages, which are specially marked articles having as title a surface form, typically followed by the word "disambiguation" (e.g., "Texas (disambiguation)"), and containing a list of references to pages for entities that are typically mentioned using that surface form.

Additionally, we extract all the surface forms used at least in two articles to refer to a Wikipedia entity page. Illustratively, the article for Pam Long contains the following *Wikitext*, which uses the surface form "Texas" to refer to Texas (TV series):

After graduation, she went to [[New York City]] and played Ashley Linden on [[Texas (TV series)|Texas]] from [[1981]] to [[1982]].

In Wikitext, the references to other Wikipedia articles are within pairs of double square brackets. If a reference contains a vertical bar then the text at the left of the bar is the name of the referred article (e.g. "Texas (TV Series)"), while the text at the right of the bar (e.g., "Texas") is the surface form that is displayed (also referred to as the anchor text of the link). Otherwise, the surface form shown in the text is identical to the title of the Wikipedia article referred (e.g., "New York City").

Using these four sources, we extracted more than 1.4 million entities, with an average of 2.4 surface forms per entity. We obtained 377k entities with one surface form, 166k entities with two surface forms, and 79k entities with three surface forms. At the other extreme, we extracted one entity with no less than 99 surface forms.

3.2 Category Information

All articles that are titled "List of […]" or "Table of […]" are treated separately as *list pages*. They were built by Wikipedia contributors to group entities of the same type together (e.g., "List of anthropologists", "List of animated television series", etc.) and are used by our system to extract category tags for the entities listed in these articles. The tags are named after the title of the Wikipedia list page. For example, from the article "List of band name etymologies", the system extracts the category tag *LIST_band_name_etymologies* and labels all the entities referenced in the list, including Texas (band), with this tag. This process resulted in the extraction of more than 1 million (entity, tag) pairs. After a post-processing phase that discards temporal tags, as well as several types of non-useful tags such as "people by name" and "places by name", we obtained a filtered list of 540 thousand pairs.

We also exploit the fact that Wikipedia enables contributors to assign *categories* to each article, which are defined as "major topics that are likely to be useful to someone reading the article". Because any Wikipedia contributor can add a category to any article and the work of filtering out bogus assignments is tedious, these categories seem to be noisier than the lists, but they can still provide a tremendous amount of information. We extracted the categories of each entity page and assigned them as tags to the corresponding entity. Again, we employed some basic filtering to discard meta-categories (e.g., "Articles with unsourced statements") and categories not useful for the process of disambiguation through tag agreement (e.g., "Living people", "1929 births"). This extraction process resulted in 2.65 million (entity, tag) pairs over a space of 139,029 category tags.

We also attempted to extract category tags based on lexicosyntactic patterns, more specifically from enumerations of entities. For example, the paragraph titled "Music of Scotland" (shown below in Wikitext) in the Wikipedia article on Scotland contains an enumeration of entities, which can be labeled *ENUM_Scotland_PAR_Music_of_Scotland*:

Modern Scottish [[pop music]] has produced many international bands including the [[Bay City Rollers]], [[Primal Scream]], [[Simple Minds]], [[The Proclaimers]], [[Deacon Blue]], [[Texas (band)|Texas]], [[Franz Ferdinand]], [[Belle and Sebastian]], and [[Travis (band)|Travis]], as well as individual artists such as [[Gerry Rafferty]], [[Lulu]], [[Annie Lennox]] and [[Lloyd Cole]], and world-famous Gaelic groups such as [[Runrig]] and [[Capercaillie (band)|Capercaillie]].

Lexicosyntactic patterns have been employed successfully in the past (e.g., Hearst, 1992; Roark and Charniak, 1998; Cederberg and Widdows, 2003), and this type of tag extraction is still a promising direction for the future. However, the brute force approach we tried – of indiscriminately tagging the entities of enumerations of four or more entities – was found to introduce a large amount of noise into the system in our development experiments.

3.3 Contexts

To extract contextual clues for an entity, we use the information present in that entity's page and in the other articles that explicitly refer to that entity.

First, the appositives in the titles of entity pages, which are eliminated to derive entity surface forms (as discussed in Section 3.1) are saved as contextual clues. For example, "TV series" becomes a context for the entity Texas (TV series).

We then extract all the entity references in the entity page. For example, from the article on Texas (band), for which a snippet in Wikitext is shown below, we extract as contexts the references pop music, Glasgow, Scotland, and so on:

'''Texas''' is a [[pop music]] band from [[Glasgow]], [[Scotland]], [[United Kingdom]]. They were founded by [[Johnny McElhone]] in [[1986 in music|1986]] and had their performing debut in [[March]] [[1988]] at […]

Reciprocally, we also extract from the same article that the entity Texas (band) is a good context for pop music, Glasgow, Scotland, etc.

The number of contexts extracted in this manner is overwhelming and had to be reduced to a manageable size. In our development experiments, we explored various ways of reducing the context information, for example, by extracting only entities with a certain number of mentions in an article, or by discarding mentions with low TF*IDF scores (Salton, 1989). In the end, we chose a strategy in which we employ as contexts for an entity two category of references: those mentioned in the first paragraph of the targeted entity page, and those for which the corresponding pages refer back to the targeted entity. For example, Pam Long and Texas (TV series) are extracted as relevant contexts for each other because their corresponding Wikipedia articles reference one another – a relevant snippet from the Pam Long article is cited in Section 3.1 and a snippet from the article for Texas (TV series) that references Pam Long is shown below:

In 1982 [[Gail Kobe]] became executive producer and [[Pam Long]] became headwriter.

In this manner, we extracted approximately 38 million (entity, context) pairs.

4 Document Analysis

In this section, we describe concisely the main text processing and entity identification components of the implemented system. We will then focus on the novel entity disambiguation component, which we propose and evaluate in this paper, in Section 5.

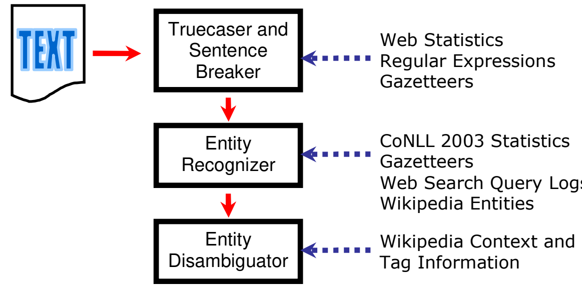

Figure 2. An overview of the processes employed by the proposed system.

Figure 2 outlines the processes and the resources that are employed by the implemented system in the analysis of text documents. First, the system splits a document into sentences and truecases the beginning of each sentence, hypothesizing whether the first word is part of an entity or it is capitalized because of orthographic conventions. It also identifies titles and hypothesizes the correct case for all words in the titles. This is done based on statistics extracted from a one-billion-word corpus, with back-off to Web statistics.

In a second stage, a hybrid named-entity recognizer based on capitalization rules, Web statistics, and statistics extracted from the CoNLL 2003 shared task data (Tjong Kim Sang and De Meulder, 2003) identifies the boundaries of the entity mentions in the text and assigns each set of mentions sharing the same surface form a probability distribution over four labels: Person, Location, Organization, and Miscellaneous.[1] The named entity recognition component resolves the structural ambiguity with regard to conjunctions (e.g., "Barnes and Noble", "Lewis and Clark"), possessives (e.g., "Alice's Adventures in Wonderland", "Britain's Tony Blair"), and prepositional attachment (e.g., "Whitney Museum of American Art", "Whitney Museum in New York") by using the surface form information extracted from Wikipedia, when available, with back-off to co-occurrence counts on the Web, in a similar way to Lapata and Keller (2004). Recursively, for each ambiguous term T_0 of the form T_1 *Particle* T_2, where *Particle* is one of a possessive pronoun, a coordinative conjunction, and a preposition, optionally followed by a determiner, and the terms T_1 and T_2 are se-

[1] While the named entity labels are used only to solve in-document coreferences by the current system, as described further in this section, preliminary experiments of probabilistically labeling the Wikipedia pages show that the these labels could also be used successfully in the disambiguation process.

quences of capitalized words and particles, we send to a search engine the query “$''T_1'' \; ''T_2''$”, which forces the engine to return only documents in which the whole terms T_1 and T_2 appear. We then count the number of times the snippets of the top $N = 200$ search results returned contain the term T_0 and compare it with an empirically obtained threshold to hypothesize whether T_0 is the mention of one entity or encompasses the mentions of two entities, T_1 and T_2.

As Wacholder et al. (1997) noted, it is fairly common for one of the mentions of an entity in a document to be a long, typical surface form of that entity (e.g., “George W. Bush”), while the other mentions are shorter surface forms (e.g., “Bush”). Therefore, before attempting to solve the semantic ambiguity, the system hypothesizes in-document coreferences and maps short surface forms to longer surface forms with the same dominant label (for example, “Brown”/$_{\text{PERSON}}$ can be mapped to “Michael Brown”/$_{\text{PERSON}}$). Acronyms are also resolved in a similar manner when possible.

In the third stage, the contextual and category information extracted from Wikipedia is used to disambiguate the entities in the text. This stage is discussed formally in Section 5 and evaluated in Section 6. Note that the performance of the disambiguation component is meaningful only when most named entity mentions are accurately identified in text. Thus, we first measured the performance of the named entity recognition component on the CoNLL 2003 test set and obtained a competitive F-measure of 0.835 (82.2% precision and 84.8% recall).

Finally, the implemented system creates hyperlinks to the appropriate pages in Wikipedia. Figure 3 shows the output of the implemented system on a sample news story, in which the identified and disambiguated surface forms are hyperlinked to Wikipedia articles.

5 The Disambiguation Component

The disambiguation process employs a vector space model, in which a vectorial representation of the processed document is compared with the vectorial representations of the Wikipedia entities.

Once the named entity surface forms were identified and the in-document coreferences hypothesized, the system retrieves all possible entity disambiguations of each surface form. Their Wikipedia contexts that occur in the document and their category tags are aggregated into a *document vector*, which is subsequently compared with the Wikipedia *entity vector* (of categories and contexts) of each possible entity disambiguation. We then choose the assignment of entities to surface forms that maximizes the similarity between the document vector and the entity vectors, as we explain further.

Formally, let $C = \{c_1,\ldots,c_M\}$ be the set of known contexts from Wikipedia and $\mathcal{T} = \{t_1,\ldots,t_N\}$ the set of known category tags. An entity e can then be represented as a vector $\delta_e \in \{0,1\}^{M+N}$, with two components, $\delta_e|_C \in \{0,1\}^M$ and $\delta_e|_{\mathcal{T}} \in \{0,1\}^N$, corresponding to the context and category information, respectively:

$$\delta_e^{\,i} = \begin{cases} 1, & \text{if } c_i \text{ is a context for entity } e \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_e^{\,M+j} = \begin{cases} 1, & \text{if } t_j \text{ is a category tag for } e \\ 0, & \text{otherwise.} \end{cases}$$

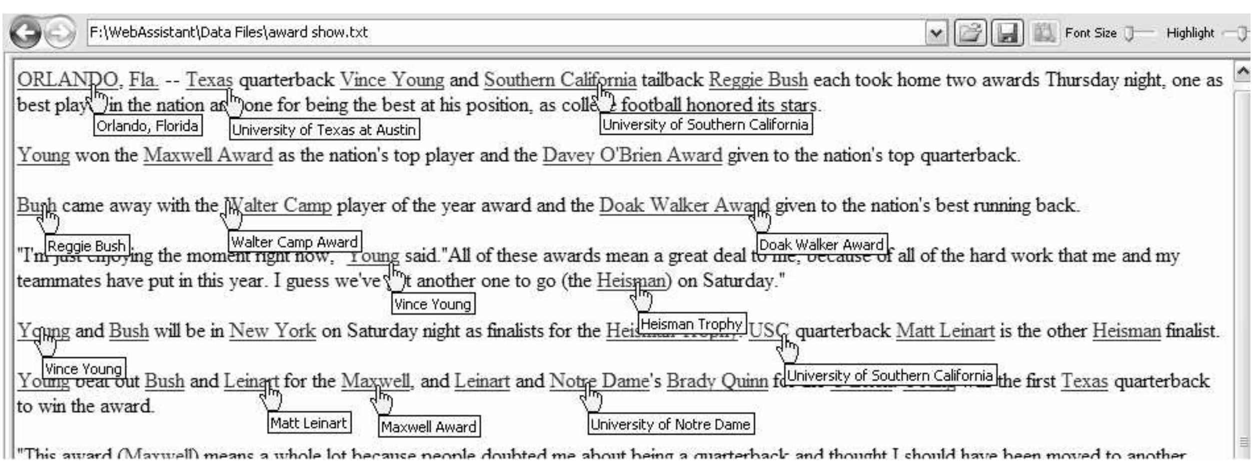

Figure 3. Screenshot of the implemented system showing an example of analyzed text. The superimposed tooltips show how several of the surface forms were disambiguated based on the context and category agreement method.

Let ε(s) denote the set of entities that are known to have a surface form s. For example, recalling Figure 1, Colombia (the country) and Columbia University are entities that are known to have the surface form "Columbia". Let D be the analyzed document and $S(D) = \{s_1,\ldots,s_n\}$ the set of surface forms identified in D. We build its context vector $d = \{d_1,\ldots,d_M\} \in \mathbb{N}^M$, where d_i is the number of occurrences of context c_i in D. To account for all possible disambiguations of the surface forms in D, we also build an extended vector $\bar{d} \in \mathbb{N}^{M+N}$ so that $\bar{d}\,|_C = d$ and $\bar{d}\,|_T = \sum_{s \in S(D)} \sum_{e \in \varepsilon(s)} \delta_e\,|_T$. [2]

Our goal is to find the assignment of entities to surface forms $s_i \mapsto e_i$, $i \in 1..n$, that maximizes the agreement between $\delta_{e_i}|_C$ and d, as well as the agreement between the categories of any two entities $\delta_{e_i}|_T$ and $\delta_{e_j}|_T$. This can be written as:

$$\underset{\substack{(e_1,\ldots,e_n)\in \\ \varepsilon(s_1)\times\ldots\times\varepsilon(s_n)}}{\arg\max} \sum_{i=1}^{n} < \delta_{e_i}|_C, d > + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} < \delta_{e_i}|_T, \delta_{e_j}|_T > , \quad (1)$$

where $< \cdot,\cdot >$ denotes the scalar product of vectors. Note that the quality of an assignment of an entity to a surface form depends on all the other assignments made, which makes this a difficult optimization problem. An arguably more robust strategy to account for category agreement, which also proves to be computationally efficient, is to maximize the agreement between the categories of the assigned entity to each surface form and all possible disambiguations of the other surface forms in D. We will show that this is equivalent to computing:

$$\underset{(e_1,\ldots,e_n)\in\varepsilon(s_1)\times\ldots\times\varepsilon(s_n)}{\arg\max} \sum_{i=1}^{n} < \delta_{e_i}, \bar{d} - \delta_{e_i}|_T > \quad (2)$$

Indeed, using the definition of $\bar{d}$ and partitioning the context and category components, we can rewrite the sum in equation (2) as

$$\sum_{i=1}^{n} < \delta_{e_i}|_C, d > + \sum_{i=1}^{n} < \delta_{e_i}|_T, \bar{d}|_T - \delta_{e_i}|_T > =$$

$$\sum_{i=1}^{n} < \delta_{e_i}|_C, d > + \sum_{i=1}^{n} < \delta_{e_i}|_T, \Big(\sum_{j=1}^{n} \sum_{e \in \varepsilon(s_j)} \delta_e|_T \Big) - \delta_{e_i}|_T > =$$

$$\sum_{i=1}^{n} < \delta_{e_i}|_C, d > + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} < \delta_{e_i}|_T, \sum_{e \in \varepsilon(s_j)} \delta_e|_T > \quad \text{(q.e.d.)}$$

Note now that the maximization of the sum in (2) is equivalent to the maximization of each of its terms, which means that the computation reduces to $\underset{e_i \in \varepsilon(s_i)}{\arg\max} < \delta_{e_i}, \bar{d} - \delta_{e_i}|_T >, i \in 1..n$., or equivalently,

$$\underset{e_i \in \varepsilon(s_i)}{\arg\max} < \delta_{e_i}, \bar{d} > - \| \delta_{e_i}|_T \|^2, i \in 1..n \quad (3)$$

Our disambiguation process therefore employs two steps: first, it builds the extended document vector and second, it maximizes the scalar products in equation (3). In practice, it is not necessary to build the document vector over all contexts C, but only over the contexts of the possible entity disambiguations of the surface forms in the document.

Also note that we are not normalizing the scalar products by the norms of the vectors (which would lead to the computation of cosine similarity). In this manner, we implicitly account for the frequency with which a surface form is used to mention various entities and for the importance of these entities (important entities have longer Wikipedia articles, are mentioned more frequently in other articles, and also tend to have more category tags).

While rarely, one surface form can be used to mention two or more different entities in a document (e.g., "Supreme Court" may refer to the federal institution in one paragraph and to a state's judicial institution in another paragraph). To account for such cases, the described disambiguation process is performed iteratively for the instances of the surface forms with multiple disambiguations with similarity scores higher than an empirically determined threshold, by shrinking the context used for the disambiguation of each instance from document level to paragraph level, and if necessary, to sentence level.

6 Evaluation

We used as development data for building the described system the Wikipedia collection as of April 2, 2006 and a set of 100 news stories on a diverse range of topics. For the final evaluation, we performed data extraction from the September 11, 2006 version of the Wikipedia collection.

We evaluated the system in two ways: on a set of Wikipedia articles, by comparing the system output with the references created by human contributors, and on a set of news stories, by doing a post-hoc evaluation of the system output. The evaluation data can be downloaded from http://research.microsoft.com/users/silviu/WebAssistant/TestData.

[2] We use the notation $\bar{d}$ to emphasize that this vector contains information that was not present in the original document D.

In both settings, we computed a disambiguation baseline in the following manner: for each surface form, if there was an entity page or redirect page whose title matches exactly the surface form then we chose the corresponding entity as the baseline disambiguation; otherwise, we chose the entity most frequently mentioned in Wikipedia using that surface form.

6.1 Wikipedia Articles

We selected at random 350 Wikipedia entity pages and we discarded their content during the information extraction phase. We then performed an automatic evaluation, in which we compared the hyperlinks created by our system with the links created by the Wikipedia contributors. In an attempt to discard most of the non-named entities, we only kept for evaluation the surface forms that started with an uppercase letter. The test articles contained 5,812 such surface forms. 551 of them referenced non-existing articles (for example, the filmography section of a director contained linked mentions of all his movies although many of them did not have an associated Wikipedia page). Also, 130 of the surface forms were not used in other Wikipedia articles and therefore both the baseline and the proposed system could not hypothesize a disambiguation for them. The accuracy on the remaining 5,131 surface forms was 86.2% for the baseline system and **88.3%** for the proposed system. A McNemar test showed that the difference is not significant, the main cause being that the majority of the test surface forms were unambiguous. When restricting the test set only to the 1,668 ambiguous surface forms, the difference in accuracy between the two systems is significant at $p = 0.01$. An error analysis showed that the Wikipedia set used as gold standard contained relatively many surface forms with erroneous or out-of-date links, many of them being correctly disambiguated by the proposed system (thus, counted as errors). For example, the test page "The Gods (band)" links to Paul Newton, the painter, and Uriah Heep, which is a disambiguation page, probably because the original pages changed over time, while the proposed system correctly hypothesizes links to Paul Newton (musician) and Uriah Heep (band).

6.2 News Stories

We downloaded the top two stories in the ten MSNBC news categories (Business, U.S. Politics, Entertainment, Health, Sports, Tech & Science, Travel, TV News, U.S. News, and World News) as of January 2, 2007 and we used them as input to our system. We then performed a post-hoc evaluation of the disambiguations hypothesized for the surface forms correctly identified by the system (i.e. if the boundaries of a surface form were not identified correctly then we disregarded it).

We defined a disambiguation to be correct if it represented the best possible Wikipedia article that would satisfy a user's need for information and incorrect otherwise. For example, the article Viking program is judged as correct for "Viking Landers", for which there is no separate article in the Wikipedia collection. Linking a surface form to a wrong Wikipedia article was counted as an error regardless whether or not an appropriate Wikipedia article existed. When the system could not disambiguate a surface form (e.g. "N' Sync", "'Bama", and "Harris County Jail"), we performed a search in Wikipedia for the appropriate entity. If an article for that entity existed (e.g., 'N Sync and Alabama) then we counted that instance as an error. Otherwise, we counted it separately as non-recallable (e.g. there is no Wikipedia article for the Harris County Jail entity and the article for Harris County, Texas does not discuss the jail system).

The test set contained 756 surface forms, of which 127 were non-recallable. The proposed system obtained an accuracy of **91.4%**, versus a 51.7% baseline (significant at $p = 0.01$). An analysis of these data showed not only that the most common surface forms used in news are highly ambiguous but also that a large number of Wikipedia pages with titles that are popular surface forms in news discuss subjects different from those with common news usage (e.g., the page titled "China" discusses the Chinese civilization and is not the correct assignment for the People's Republic of China entity; similarly, the default page for "Blackberry" talks about the fruit rather than the wireless company with the same name).

7 Conclusions and Potential Impact

We presented a large scale named entity disambiguation system that employs a huge amount of information automatically extracted from Wikipedia over a space of more than 1.4 million entities. In tests on both real news data and Wikipedia text, the system obtained accuracies exceeding 91% and 88%. Because the entity recognition and disambiguation processes employed use very little language-dependent resources additional to Wikipedia, the system can be easily adapted to languages other than English.

The system described in this paper has been fully implemented as a Web browser (Figure 3), which can analyze any Web page or client text document. The application on a large scale of such an entity extraction and disambiguation system could result in a move from the current space of words to a space of concepts, which enables several paradigm shifts and opens new research directions, which we are currently investigating, from entity-based indexing and searching of document collections to personalized views of the Web through entity-based user bookmarks.

Acknowledgments

The author would like to gratefully thank Mike Schultz and Robert Ragno for their help in building the system infrastructure, Microsoft Live Search for providing access to their search engine, and the anonymous reviewers for their useful comments.

References

Bagga, A. and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of COLING-ACL*, 79-85.

Bunescu, R. and M. Paşca. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of EACL*, 9-16.

Cederberg, S. and D. Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of hyponymy extraction. In *Proceedings of CoNLL*, 111-118.

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. ACE program – task definitions and performance measures. In *Proceedings of LREC*, 837-840.

Edmonds, P. and S. Cotton. 2001. Senseval-2 overview. In *Proceedings of SENSEVAL-2*, 1-6.

Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of IJCAI*, 1606-1611.

Gale, W., K. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA SNL Workshop*, 233-237.

Grishman, R. and B. Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proceedings of COLING*, 466-471.

Hearst, M. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. COLING*, 539-545.

Hirschman, L. and N. Chinchor. 1997. MUC-7 Coreference Task Definition. In *Proceedings of MUC-7*.

Kanada, Y. 1999. A method of geographical name extraction from Japanese text. In *Proceedings of CIKM*, 46-54.

Kilgarriff, A. and J. Rosenzweig. 2000. Framework and results for English Senseval. *Computers and Humanities, Special Issue on SENSEVAL*, 15-48.

Lapata, M. and F. Keller. 2004. The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks. In *Proceedings of HLT,* 121-128.

Mann, G. S. and D. Yarowsky. 2003. Unsupervised Personal Name Disambiguation. In *Proceedings of CoNLL*, 33-40.

Mihalcea, R., T. Chklovski, and A. Kilgarriff. The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3*, 25-28.

Overell, S., and S. Rüger. 2006 Identifying and grounding descriptions of places. In *SIGIR Workshop on Geographic Information Retrieval*.

Raghavan, H., J. Allan, and A. McCallum. 2004. An exploration of entity models, collective classification and relation description. In *KDD Workshop on Link Analysis and Group Detection*.

Ravin, Y. and Z. Kazi. 1999. Is Hillary Rodham Clinton the President? In *ACL Workshop on Coreference and it's Applications*.

Remy, M. 2002. Wikipedia: The free encyclopedia. In *Online Information Review*, 26(6): 434.

Roark, B. and E. Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of COLING-ACL*, 1110-1116.

Salton, G. 1989. *Automatic Text Processing*. Addison-Wesley.

Smith, D. A. and G. Crane. 2002. Disambiguating geographic names in a historic digital library. In *Proceedings of ECDL*, 127-136.

Strube, M. and S. P. Ponzeto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of AAAI*, 1419-1424.

Tjong Kim Sang, E. F. and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL*, 142-147.

Wacholder, N., Y. Ravin, and M. Choi. 1997. Disambiguation of proper names in text. In *Proceedings of ANLP*, 202-208.

Woodruff, A. G. and C. Paunt. GIPSY:Automatic geographic indexing of documents. *Journal of the American Society for Information Science and Technology*, 45(9):645-655.