Towards Robust Unsupervised Personal Name Disambiguation

Ying Chen
Center for Spoken Language Research
University of Colorado at Boulder
yc@colorado.edu

James Martin
Department of Computer Science
University of Colorado at Boulder
James.Martin@colorado.edu

Abstract

The increasing use of large open-domain document sources is exacerbating the problem of ambiguity in named entities. This paper explores the use of a range of syntactic and semantic features in unsupervised clustering of documents that result from ad hoc queries containing names. From these experiments, we find that the use of robust syntactic and semantic features can significantly improve the state of the art for disambiguation performance for personal names for both Chinese and English.

1 Introduction

An ever-increasing number of question answering, summarization and information extraction systems are coming to rely on heterogeneous sets of documents returned by open-domain search engines from collections over which application developers have no control. A frequent special case of these applications involves queries containing named entities of various sorts and receives as a result a large set of possibly relevant documents upon which further deeper processing is focused. Not surprisingly, many, if not most, of the returned documents will be irrelevant to the goals of the application because of the massive ambiguity associated with the query names of people, places and organizations in large open collections. Without some means of separating documents that contain mentions of distinct entities, most of these applications will produce incorrect results. The work presented here, therefore, addresses the problem of automatically separating sets of news documents generated by queries containing personal names into coherent partitions.

The approach we present here combines unsupervised clustering methods with robust syntactic and semantic processing to automatically cluster returned news documents (and thereby entities) into homogeneous sets. This work follows on the work of Bagga & Baldwin (1998), Mann & Yarowsky (2003), Niu et al. (2004), Li et al. (2004), Pedersen et al. (2005), and Malin (2005). The results described here advance this work through the use of syntactic and semantic features that can be robustly extracted from the kind of arbitrary news texts typically returned from open-domain sources.

The specific contributions reported here fall into two general areas related to robustness. In the first, we explore the use of features extracted from syntactic and semantic processing at a level that is robust to changes in genre and language. In particular, we seek to go beyond the kind of bag of local words features employed in earlier systems (Bagga & Baldwin, 1998; Gooi & Allan, 2004; Pedersen et al., 2005) that did not attempt to exploit deep semantic features that are difficult to extract, and to go beyond the kind of biographical information (Mann & Yarowsky, 2003) that is unlikely to occur with great frequency (such as place of birth, or family relationships) in many of the documents returned by typical search engines. The second contribution involves the application of these techniques to both English and Chinese news collections. As we'll see, the methods are effective with both, but error analyses reveal interesting differences between the two languages.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 190–198, Prague, June 2007.

The paper is organized as follows. Section 2 addresses related work and compares our work with that of others. Section 3 introduces our new phrase-based features along two dimensions: from syntax to semantics; and from local sentential contexts to document-level contexts. Section 4 first describes our datasets and then analyzes the performances of our system for both English and Chinese. Finally, we draw some conclusions.

2 Previous work

Personal name disambiguation is a difficult problem that has received less attention than those topics that can be addressed via supervised learning systems. Most previous work (Bagga & Baldwin, 1998; Mann & Yarowsky, 2003; Li et al., 2004; Gooi & Allan, 2004; Malin, 2005; Pedersen et al., 2005; Byung-Won On and Dongwon Lee, 2007) employed unsupervised methods because no large annotated corpus is available and because of the variety of the data distributions for different ambiguous personal names.

Since it is common for a single document to contain one or more mentions of the ambiguous personal name of interest, there is a need to define the object to be disambiguated (the ambiguous object). In Bagga & Baldwin (1998), Mann & Yarowsky (2003) and Gooi & Allan (2004), an ambiguous object refers to a single entity with the ambiguous personal name in a given document. The underlying assumption for this definition is "one person per document" (all mentions of the ambiguous personal name in one document refer to the same personal entity in reality). In Niu et al. (2004) and Pedersen et al. (2005), an ambiguous object is defined as a mention of the ambiguous personal name in a corpus.

The first definition of the ambiguous object (document-level object) can include much information derived from that document, so that it can be represented by rich features. The later definition of the ambiguous object (mention-level object) can simplify the detection of the ambiguous object, but because of the limited coverage, it usually can use only local context (the text around the mention of the ambiguous personal name) and might miss some document-level information. The kind of name disambiguation based on mention-level objects really solves "within-document name ambiguity" and "cross-document name ambiguity" simultaneously, and often has a higher performance than the kind based on document-level objects because two mentions of the ambiguous personal name in a document are very likely to refer to the same personal entity. From our news corpus, we also found that mentions of the ambiguous personal name of interest in a news article rarely refer to multiple entities, so our system will focus on the name disambiguation for document-level objects.

In general, there are two types of information usually used in name disambiguation (Malin, 2005): personal information and relational information (explicit and implicit). Personal information gives biographical information about the ambiguous object, and relational information specifies explicit or implicit relations between the ambiguous object and other entities, such as a membership relation between "John Smith" and "Labor Party." Usually, explicit relational information can be extracted from local context, and implicit relational information is far away from the mentions of the ambiguous object. A hard case of name disambiguation often needs implicit relational information that provides a background for the ambiguous object. For example, if two news articles in consideration report an event happening in "Labor Party," this implicit relational information between "John Smith" and "Labor Party" can give a hint for name disambiguation if no personal or explicit relational information is available.

Bagga & Baldwin (1998), Mann & Yarowsky (2003), Gooi & Allan (2004), Niu et al. (2004), and Pedersen et al. (2005) explore features in local context. Bagga & Baldwin (1998), Gooi & Allan (2004), and Pedersen et al. (2005) use local token features; Mann & Yarowsky (2003) extract local biographical information; Niu et al. (2004) use co-occurring Named Entity (NE) phrases and NE relationships in local context. Most of these local contextual features are personal information or explicit relational information.

Li et al. (2004) and Malin (2005) consider named-entity disambiguation as a graph problem, and try to capture information related to the ambiguous object beyond local context, even implicit relational information. Li et al. (2004) use the EM algorithm to learn the global probability distribution among documents, entities, and representative mentions, and Malin (2005) constructs a social network graph to learn a similarity matrix.

In this paper, we also explore both personal and relational information beyond local context. But we achieve it with a different approach: extracting these types of information by means of syntactic and semantic processing. We not only extract local NE phrases as in Niu et al. (2004), but also use our entity co-reference system to extract accurate and representative NE phrases occurring in a document which may have a relation to the ambiguous object. At the same time, syntactic phrase information sometimes can overcome the imperfection of our NE system and therefore makes our disambiguation system more robust.

3 Overall Methodology

Our approach follows a common architecture for named-entity disambiguation: the detection of ambiguous objects, feature extraction and representation, similarity matrix learning, and finally clustering.

In our approach, all documents are preprocessed with a syntactic phrase chunker (Hacioglu, 2004) and the EXERT[1] system (Hacioglu et al. 2005; Chen & Hacioglu, 2006), a named-entity detection and co-reference resolution system that was developed for the ACE[2] project. A syntactic phrase chunker segments a sentence into a sequence of base phrases. A base phrase is a syntactic-level phrase that does not overlap another base phrase. Given a document, the EXERT system first detects all mentions of entities occurring in that document (named-entity detection) and then resolves the different mentions of an entity into one group that uniquely represents the entity (within-document co-reference resolution). The ACE 2005 task can detect seven types of named entities: person, organization, geo-political entity, location, facility, vehicle, and weapon; each type of named entity can occur in a document with any of three distinct formats: name, nominal construction, and pronoun. The F score of the syntactic phrase chunker, which is trained and tested on the Penn TreeBank, is 94.5, and the performances of the EXERT system are 82.9 (ACE value for named-entity detection) and 68.5 (ACE value for within-document co-reference resolution).

3.1 The detection of ambiguous objects

In our approach, we assume that the ambiguous personal name has already been determined by the application. Moreover, we adopt the policy of "one person per document" as in Bagga & Baldwin (1998), and define an ambiguous object as a set of target entities given by the EXERT system. A target entity is an entity that has a mention of the ambiguous personal name. Given the definition of an ambiguous object, we define a local sentence (or local context) as a sentence that contains a mention of any target entity.

3.2 Feature extraction and representation

Since considerable personal and relational information related to the ambiguous object resides in the noun phrases in the document, such as the person's job and the person's location, we attempt to capture this noun phrase information along two dimensions: from syntax to semantics, and from local contexts to document-level contexts.

Base noun phrase feature: To keep this feature focused, we extract only noun phrases occurring in the local sentences and the summarized sentences (the headline + the first sentence of the document) of the document. The local sentences usually include personal or explicit relational information about the ambiguous object, and the summarized sentences of a news document usually give a short summary of the whole news story. With the syntactic phrase chunker, we develop two base noun phrase models: (i) ***Contextual base noun phrases (Contextual bnp)***, the base noun phrases in the local sentences; (ii) ***Summarized base noun phrases (Summarized bnp)***, the base noun phrases in the local sentences and the summarized sentences. A base noun phrase of interest serves as an element in the feature vector.

Named-Entity feature: Given the EXERT system, a direct and simple way to extract some semantic information is to use the named entities detected in the document. Based on their relationship to the ambiguous personal name, the named entities identified in a text can be divided into three categories:

(i) Target entity: an entity that has a mention of the ambiguous personal name. Target entities often include some personal information about the ambiguous object, such as the title, position, and so on.

[1] http://sds.colorado.edu/EXERT
[2] http://projects.ldc.upenn.edu/ace/

Feature Space

Contextual base noun phrases' feature vector: < Hope Mills police Capt. John Smith[16], what[16], he[16], the statements[16], no criminal violation[16], what[17], the individuals[17], no direct threat[17], Smith[17], He and Thomas[18], they[18], Collins[18], his bill[18]>

Summarized base noun phrases' feature vector: < Hope Mills police Capt. John Smith[16], what[16], he[16], the statements[16], no criminal violation[16], what[17], the individuals[17], no direct threat[17], Smith[17], He and Thomas[18], they[18], Collins[18], his bill[18], Collins[1], restaurant[1], HOPE MILLS[2], Commissioner Tonzie Collins[2], a town restaurant[2], an alleged run-in[2], two workers[2], Feb. 21[2]>

Contextual entities' feature vector: **<** Hope Mills police Capt. John Smith[16], Jenny Thomas[4], Commissioner Tonzie Collins[2], He and Thomas[4], the individuals[17]**>**

Document entities' feature vector: **<** Hope Mills police Capt. John Smith [16], Jenny Thomas[4], Commissioner Tonzie Collins[2], He and Thomas[4], the individuals[17], Andy's Cheesesteaks[4], HOPE MILLS [2], two workers[2], the Village Shopping Center [4], Hope Mills Road [4] **>**

Entity space

Target entity: < Hope Mills police Capt. John Smith[16], he[16], Smith[17], He[18]>
Local entity: < Thomas[18], Jenny Thomas[4], manager[4]>,
< Collins[18], his[18], Collins[1], Commissioner Tonzie Collins [2]>, ……
Non-local entity: < restaurant[1], a town restaurant[2], there[2], Andy's Cheesesteaks[4]>, ……

Text space

(Headline & S1) Collins banned from restaurant
(S2) HOPE MILLS — Commissioner Tonzie Collins has been banned from a town restaurant after an alleged run-in with two workers there Feb. 21. ……
(S4) "In all fairness, that is not a representation of the town," said Jenny Thomas, manager at Andy's Cheesesteaks in the Village Shopping Center on Hope Mills Road. ……
(S16) **Hope Mills police Capt. John Smith** said based on what **he** read in the statements, no criminal violation was committed.
(S17) "Based on what the individuals involved said, there was no direct threat," **Smith** said.
(S18) **He** and Thomas said they don't think Collins intentionally left without paying his bill. ……

Figure 1: A Sample of Feature Extraction

(ii) Local entity: an entity other than a target entity that has a mention occurring in any local sentence. Local entities often include entities that are closely related to the ambiguous object, such as employer, location and co-workers.

(iii) Non-local entity: an entity that is not either the local entity or the target entity. Non-local entities are often implicitly related to the ambiguous object and provide background information for the ambiguous object.

To assess how important these entities are to named-entity disambiguation, we create two kinds of entity models: (i) ***Contextual entities:*** the entities in the feature vector are target entities and local entities; (ii) ***Document entities:*** the entities in the feature vector include all entities in the document including target entities, local entities and non-local entities.

Since a given entity can be represented by many mentions in a document, we choose a single representative mention to represent each entity. The representative mention is selected according to the following ordered preference list: longest NAME mention, longest NOMINAL mention. A representative mention phrase serves as an element in a feature vector.

Although the mentions of contextual entities often overlap with contextual base noun phrases, the representative mention of a contextual entity often goes beyond local sentences, and is usually the first or longest mention of that contextual entity. Compared to contextual base noun phrases, the representative mention of a contextual entity often includes more detail and accurate information about the entity. On the other hand, the contextual base noun phrase feature detects all noun phrases occurring in local sentences that are not limited to the seven types of named entities discovered by the EXERT system. Compared to the contextual entity feature, the contextual base noun phrase feature is more general and can sometimes overcome errors propagated from the named-entity system.

To make this more concrete, the feature vectors for a document containing "John Smith" are highlighted in Figure 1. The superscript number for each phrase refers to the sentence where the phrase is located, and we assume that the syntactic phrase chunker and the EXERT system work perfectly.

3.3 Similarity matrix learning

Given a pair of feature vectors consisting of phrase-based features, we need to choose a similarity scheme to calculate the similarity. Because of the word-space delimiter in English, the feature vector for an English document comprises phrases, whereas that for a Chinese document comprises tokens. There are a number of similarity schemes for learning a similarity matrix from token-based feature vectors, but there are few schemes for phrase-based feature vectors.

Cohen et al. (2003) compared various similarity schemes for the task of matching English entity names and concluded that the hybrid scheme they call SoftTFIDF performs best. SoftTFIDF is a token-based similarity scheme that combines a standard TF-IDF weighting scheme with the Jaro-Winkler distance function. Since Chinese feature vectors are token-based, we can directly use SoftTFIDF to learn the similarity matrix. However, English feature vectors are phrase-based, so we need to run SoftTFIDF iteratively and call it "two-level SoftTFIDF." First, the standard SoftTFIDF is used to calculate the similarity between phrases in the pair of feature vectors; in the second phase, we reformulate the standard SoftTFIDF to calculate the similarity for the pair of feature vectors.

First, we introduce the standard SoftTFIDF. In a pair of feature vectors S and T, $S = (s_1, \ldots, s_n)$ and $T = (t_1, \ldots, t_m)$. Here, s_i (i = 1…n) and t_j (j = 1…m) are substrings (tokens). Let $CLOSE(\theta; S;T)$ be the set of substrings $w \in S$ such that there is some $v \in T$ satisfying $dist(w; v) > \theta$. The Jaro-Winkler distance function (Winkler, 1999) is $dist(;)$. For $w \in CLOSE(\theta; S;T)$, let $D(w; T) = \max_{v \in T} dist(w; v)$. Then the standard SoftTFIDF is computed as

$$\mathrm{SoftTFIDF}(S,T) = \sum_{w \in CLOSE(\theta;S;T)} \mathrm{V}(w, S) \times \mathrm{V}(w, T) \times \mathrm{D}(w, T)$$

$$\mathrm{V}'(w, S) = \log(\mathrm{TF}_{w,S} + 1) \times \log(\mathrm{IDF}_w)$$

$$\mathrm{V}(w, S) = \frac{\mathrm{V}\ (w, S)}{\sqrt{\sum_{w \in S} \mathrm{V}(w, S)^2}}$$

where $\mathrm{TF}_{w,S}$ is the frequency of substrings w in S, and IDF_w is the inverse of the fraction of documents in the corpus that contain w. In computing the similarity for the English phrase-based feature vectors, in the second step of "two-level SoftTFIDF," the substring w is a phrase and *dist* is the standard SoftTFIDF.

So far, we have developed several feature models and learned the corresponding similarity matrices, but clustering usually needs only one unique similarity matrix. Since a feature may have different effects for the disambiguation depending on the ambiguous personal name in consideration, to achieve the best disambiguation ability, each personal name may need its own weighting scheme to combine the given similarity matrices. However, learning that kind of weighting scheme is very difficult, so in this paper, we simply combine the similarity matrices, assigning equal weight to each one.

3.4 Clustering

Although clustering is a well-studied area, a remaining research problem is to determine the optimal parameter setting during clustering, such as the number of clusters or the stop-threshold, a problem that is important for real tasks and that is not at all trivial.

Since the focus of this paper is only on feature development, we simply employ a clustering method that can reflect the quality of the similarity matrix for clustering. Here, we choose agglomerative clustering with a single linkage. Since each personal name may need a different parameter setting, to test the importance of the parameter setting for clustering, we use two kinds of stop-thresholds for agglomerative clustering in our experiments: first, to find the optimal stop-threshold for any ambiguous personal name and for each feature model, we run agglomerative clustering with all possible stop-thresholds, and choose the one that has the best performance as the optimal stop-threshold; second, we use a fixed stop-threshold acquired from development data.

4 Performance

4.1 Data

To capture the real data distribution, we use two sets of naturally occurring data: Bagga's corpus and the Boulder Name corpus, which is a news corpus locally acquired from a web search. Bagga's corpus is a document collection for the English personal name "John Smith" that was used by Bagga & Baldwin (1998). There are 256 articles that match the "/John.*?Smith/" regular expression in 1996 and 1997 editions of the *New York Times*, and 94 distinct "John Smith" personal entities are mentioned. Of these, 83 "John Smiths" are mentioned in only one article (singleton clusters containing only one object), and 11 other "John Smiths" are mentioned several times in the remaining 173 articles (non-singleton clusters containing more than one object). For the task of cross-document co-reference, Bagga & Baldwin (1998) chose 24 articles from 83 singleton clusters, and 173 other articles in 11 non-singleton clusters to create the final test data set – Bagga's corpus.

We collected the Boulder Name corpus by first selecting four highly ambiguous personal names each in English and Chinese. For each personal name, we retrieved the first non-duplicated 100 news articles from Google (Chinese) or Google news (English). There are four data sets for English personal names and four data sets for Chinese personal names: James Jones, John Smith, Michael Johnson, Robert Smith, and Li Gang, Li Hai, Liu Bo, Zhang Yong.

Compared to Bagga's corpus, which is limited to the *New York Times*, the documents in the Boulder Name corpus were collected from different sources, and hence are more heterogeneous and noisy. This variety in the Boulder Name corpus reflects the distribution of the real data and makes named-entity disambiguation harder.

For each ambiguous personal name in both corpora, the gold standard clusters have a long-tailed distribution - a high percentage of singleton clusters plus a few non-singleton clusters. For example, in the 111 documents containing "John Smith" in the Boulder Name corpus, 53 "John Smith" personal entities are mentioned. Of them, 37 "John Smiths" are mentioned only once. The long-tailed distribution brings some trouble to clustering, since in many clustering algorithms a singleton cluster is considered as a noisy point and therefore is ignored.

4.2 Corpus performance

Because of the long tail of the data set, we design a baseline using one cluster per document. To evaluate our disambiguation system, we choose the B-cubed scoring method that was used by Bagga & Baldwin (1998).

In order to compare our work to that of others, we re-implement the model used by Bagga & Baldwin (1998). First, extracting all local sentences produces a summary about the given ambiguous object. Then, the object is represented by the tokens in its summary in the format of a vector, and the tokens in the feature vector are in their morphological root form and are filtered by a stop-word dictionary. Finally, the similarity matrix is learned by the TF-IDF method.

Because both "two-level SoftTFIDF" and agglomerative clustering require a parameter setting, for each language, we reserve two ambiguous personal names from the Boulder Name corpus as development data (John Smith, Michael Johnson, Li Gang, Zhang Yong), and the other data are reserved as test data: Bagga's corpus and the other personal names in the Boulder Name corpus (Robert Smith, James Jones, Li Hai, Liu Bo).

For any ambiguous personal name and for each feature model, we find the optimal stop-threshold for agglomerative clustering, and show the corresponding performances in Table 1, Table 2 and Table 3. However, for the most robust feature model, Bagga + summarized bnp + document entities, we learn the fixed stop-threshold for agglomerative clustering from the development data (0.089 for English data and 0.078 for Chinese data), and show the corresponding performances in Table 4.

4.2.1 Performance on Bagga's corpus

Table 1 shows the performance of each feature model for Bagga's corpus with the optimal stop-threshold. The metric here is the B-cubed F score (precision/recall).

Because of the difference between Bagga's resources and ours (different versions of the named-entity system and different dictionaries of the morphological root and the stop-words), our best

B-cubed F score for Bagga's model is 80.3— 4.3 percent lower than the best performance reported by Bagga & Baldwin (1998): 84.6.

From Table 1, we found that the syntactic features (contextual bnp and summarized bnp) and semantic features (contextual entities and document entities) consistently improve the performances, and all performances outperform the best result reported by Bagga & Baldwin (1998): 84.6

Model	B-cubed performance
Gold standard cluster #	35
Baseline	30.17 (100.00/17.78)
Bagga	80.32 (94.77/69.70)
Bagga + contextual bnp	89.16 (89.18/89.13)
Bagga + summarized bnp	89.59 (92.60/86.78)
Bagga + summarized bnp + contextual entities	89.60 (87.16/92.18)
Bagga + summarized bnp + document entities	92.02 (93.10/90.97)

Table 1: Performances for Bagga's corpus with the optimal stop-threshold

Name Model	John Smith (dev)	Michael Johnson (dev)	Robert Smith (test)	James Jones (test)	Average performance
Gold standard cluster #	53	52	65	24	
Baseline	64.63 (111)	67.97 (101)	78.79 (100)	37.50 (104)	62.22
Bagga	82.63 (75)	89.07 (66)	91.56 (73)	86.42 (24)	87.42
Bagga + contextual bnp	85.18 (62)	89.13 (65)	92.35 (74)	86.45 (22)	88.28
Bagga + summarized bnp	85.97 (66)	91.08 (51)	93.17 (70)	90.11 (33)	90.08
Bagga + summarized bnp + contextual entities	85.44 (70)	94.24 (55)	91.94 (73)	96.66 (24)	92.07
Bagga + summarized bnp + document entities	91.94 (61)	92.55 (51)	93.48 (67)	97.10 (28)	93.77

Table 2: Performances for the English Boulder Name corpus with the optimal stop-threshold

Name Model	Li Gang (dev)	Zhang Yong (dev)	Li Hai (test)	Liu Bo (test)	Average performance
Gold standard cluster #	57	63	57	45	
Baseline	72.61 (100)	76.83 (101)	74.03 (97)	62.07 (100)	71.39
Bagga	96.21 (57)	96.43 (64)	94.51 (64)	91.66 (49)	94.70
Bagga + contextual bnp	97.57 (57)	96.38 (66)	94.53 (64)	93.21 (51)	95.42
Bagga + summarized bnp	98.50 (56)	96.17 (61)	95.38 (62)	93.21 (51)	95.81
Bagga + summarized bnp + contextual entities	99.50 (58)	95.49 (63)	96.75 (58)	91.05 (52)	95.70
Bagga + summarized bnp + document entities	99.50 (56)	94.57 (70)	98.57 (59)	97.02 (48)	97.41

Table 3: Performances for the Chinese Boulder Name corpus with the optimal stop-threshold

English Name	John Smith (dev)	Michael Johnson (dev)	Robert Smith (test)	James Jones (test)	Average performance
Bagga + summarized bnp + document entities	91.31 (*91.94*)	90.57 (*92.55*)	86.71 (*93.48*)	96.64 (*97.10*)	91.31 (*93.77*)
Chinese Name	Li Gang (dev)	Zhang Yong (dev)	Li Hai (test)	Liu Bo (test)	Average performance
Bagga + summarized bnp + document entities	99.06 (*99.50*)	94.56 (*94.56*)	98.25 (*98.57*)	89.18 (*97.02*)	95.26 (*97.41*)

Table 4: Performances for the Boulder Name corpus with the fixed stop-threshold

4.2.2 Performance on the Boulder Name corpus

Table 2 and Table 3 show the performance of each feature model with the optimal stop-threshold for the English and Chinese Boulder Name corpora, respectively. The metric is the B-cubed F score and the number in brackets is the corresponding cluster number. Since the same feature model has different contributions for different ambiguous personal names, we list the average performances for all ambiguous names in the last column in both tables.

Comparing Table 2 and Table 3, we find that Bagga's model has different performances for the English and Chinese corpora. That means that contextual tokens have different contributions in the two languages. There are three apparent causes for this phenomenon. The first concerns the frequency of pronouns in English vs. pro-drop in Chinese. The typical usage of pronouns in English requires an accurate pronoun co-reference resolution that is very important for the local sentence extraction in Bagga's model. In the Boulder Name corpus, we found that ambiguous personal names occur in Chinese much more frequently than in English. For example, the string "Liu Bo" occurs 876 times in the "Liu Bo" data, but the string "John Smith" occurs only 161 times in the "John Smith" data. The repetition of ambiguous personal names in Chinese reduces the burden on pronoun co-reference resolution and hence captures local information more accurately.

The second factor is the fact that tokens in Bagga's model for Chinese are words, but a Chinese word is a unit bigger than an English word, and may contain more knowledge. For example, "the White House" has three words in English, and a word in Chinese. Since Chinese named-entity detection can be considered a sub-problem of Chinese word segmentation, a word in Chinese can catch partial information about named entities.

Finally, compared to Chinese news stories, English news stories are more likely to mention persons marginal to the story, and less likely to give the complete identifying information about them in local context. Those phenomena require more background information or implicit relational information to improve English named-entity disambiguation.

From Table 2 and Table 3, we see that the average performance of all ambiguous personal names is increased (from 87.42 to 93.77 for English and from 94.70 to 97.41 for Chinese) by incorporating more information: contextual bnp (contextual base noun phrases), summarized bnp (summarized base noun phrases), contextual entities, and document entities. This indicates that the phrase-based features, the syntactic and semantic noun phrases, are very useful for disambiguation.

From Table 2 and Table 3, we also see that the phrase-based features can improve the average performance, but not always for all ambiguous personal names. For example, the feature model "Bagga + summarized bnp + contextual entities" hurts the performance for "Robert Smith." As we mentioned above, the Boulder Name corpus is heterogeneous, so each feature does not make the same contribution to the disambiguation for any ambiguous personal name. What we need to do is to find a feature model that is robust for all ambiguous personal names.

In Table 4, we choose the last feature model—Bagga + summarized bnp + document entities—as the final feature model, learn the fixed stop-threshold for clustering from the development data, and show the corresponding performances as B-cubed F scores. The performances in italics are the performances with the optimal stop-threshold. From Table 4, we find that, with the exception of "Robert Smith" and "Liu Bo," the performances for other ambiguous personal names with the fixed threshold are close to the corresponding best performances.

5 Conclusion

This work has explored the problem of personal named-entity disambiguation for news corpora. Our experiments extend token-based information to include noun phrase-based information along two dimensions: from syntax to semantics, and from local sentential contexts to document-level contexts. From these experiments, we find that rich and broad information improves the disambiguation performance considerably for both English and Chinese. In the future, we will continue to explore additional semantic features that can be robustly extracted, including features derived from semantic relations and semantic role labels, and try to extend our work from news articles to web pages that include more noisy information. Finally, we have focused here primarily on feature development and not on clustering. We believe that the skewed long-tailed distribution that characterizes this data requires the use of clustering algorithms tailored to this distribution. In particular, the large number of singleton clusters is an issue that confounds the standard clustering methods we have been employing.

References

A. Bagga and B. Baldwin. 1998. *Entity–based Cross–document Co–referencing Using the Vector Space Model*. In 17th COLING.

Y. Chen and K. Hacioglu. 2006. *Exploration of Coreference Resolution: The ACE Entity Detection and Recognition Task*. In 9th International Conference on TEXT, SPEECH and DIALOGUE.

W. Cohen, P. Ravikumar, S. Fienberg. 2003. *A Comparison of String Metrics for Name-Matching Tasks*. In IJCAI-03 II-Web Workshop.

C. H. Gooi and J. Allan. 2004. *Cross-Document Coreference on a Large Scale Corpus*. In NAACL

K. Hacioglu, B. Douglas and Y. Chen. 2005. *Detection of Entity Mentions Occurring in English and Chinese Text*. Computational Linguistics.

K. Hacioglu. 2004. *A Lightweight Semantic Chunking Model Based On Tagging*. In HLT/NAACL.

X. Li, P. Morie, and D. Roth. 2004. *Robust Reading: Identification and Tracing of Ambiguous Names*. In Proc. of NAACL, pp. 17—24.

B. Malin. 2005. *Unsupervised Name Disambiguation via Social Network Similarity*. SIAM.

G. Mann and D. Yarowsky. 2003. *Unsupervised Personal Name Disambiguation*. In Proc. of CoNLL-2003, Edmonton, Canada.

C. Niu, W. Li, and R. K. Srihari. 2004. *Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction.* In ACL

B. On and D. Lee. 2007. *Scalable Name Disambiguation using Multi-Level Graph Partition*. SIAM.

T. Pedersen, A. Purandare and A. Kulkarni. 2005. *Name Discrimination by Clustering Similar Contexts*. In Proc. of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics, pages 226-237. Mexico City, Mexico.

T. Pedersen and A. Kulkarni. 2007. *Unsupervised Discrimination of Person Names in Web Contexts*. In Proc. of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics.

W. E. Winkler. 1999. *The state of record linkage and current research problems. Statistics of Income Division*, Internal Revenue Service Publication R99/04.

A. Yates and O. Etzioni. 2007. *Unsupervised Resolution of Objects and Relations on the Web*. In NAACL.