A Personalized Approach to Named Entity Disambiguation:
Exploiting user-specific structured knowledge to model author interests and resolve intended meaning

Abstract:

Recent years have seen an explosion in the amount of user-generated content found on the web, presenting both new challenges and new opportunities in information retrieval. This paper focuses on those associated with the disambiguation of named entities in short texts, and we introduce a personalized approach to this problem that is novel in its incorporation of user-specific, structured external information.

A user's text based content on social media sites often contains many mentions of people, places, and events. Unfortunately, this content is often noisy and hard to process, making it difficult to identify the intended referent when an entity mention is ambiguous. Further, in the case of short texts such as annotations, tags, or micro-blog posts there is little lexical context on which traditional disambiguation strategies rely.

Opportunely, at the same time, an individual often participates in multiple online communities, and a user's maintained accounts specify profile information that can serve to uniquely identify her across different websites. Also, a user's contributions to knowledge repositories offer a source of structured data indicative of personal areas of interest and topics of attention.

We aim to uncover, connect, and leverage this external information as a kind of "personal context" that can be applied to improve existing disambiguation strategies. By modeling a user's domains of interest with respect to a structured knowledge base, we reduce the search space of candidate entity meanings to those topics about which a user has previously exhibited interest and thereby achieve better accuracy during entity resolution tasks.

1 INTRODUCTION & BACKGROUND

Named Entity Recognition (NER) refers to the process of systematically identifying in unstructured text mentions of entities such as people, places, or organizations. A major associated challenge is the *Named Entity Disambiguation (NED)* problem, which occurs when it is unclear to what actual entity a detected entity's *surface form* may refer. For example, the *surface* "Ford" is an ambiguous named entity whose mention could refer to a number of candidate *meanings* or *concepts*, of which there are over 50 according to its Wikipedia disambiguation page.

Typical disambiguation strategies operate by constructing a representation for an entity mention's context and that mention's candidate meanings, formulating a measure of similarity between the context and each meaning, and selecting as the

correct candidate the one with the greatest computed similarity score. However, entities in short texts on social media sites are inherently difficult to disambiguate due to the nature of this content. In particular, lexical context is sparse or non-existent in short texts. In addition, all of a user's contributions on a site cannot serve as a feasible corpus or as background information since users do not generally produce large enough volumes of such content. It therefore becomes desirable to draw on external data.

Wikipedia is a rich source of external information, providing broad coverage of domain independent named entity topics and rare word senses (Zesch, 2007; Li, 2011). Separate research has also revealed that it is possible to perform topic modeling on the text a user produces online in order to model individual interests. We propose to combine these two ideas and apply them to the problem of Named Entity Disambiguation.

Specifically, we propose that the article revision history of a Wikipedia editor can serve as a basis on which to model the topics and concepts that most interest a user and about which that user tends to produce content as a result. We also hypothesize that such individual interests are expressed across the multiple online communities in which the user participates. Thus, we aim to connect the identity of the author of an ambiguous text with the Wikipedia editor account that belongs to the same individual. By measuring the overlap between the concepts associated with a user's interest model and the concept that corresponds to a potential meaning of an ambiguous entity, we are able to take advantage of personally and semantically relevant information in order to determine the user's actual intended meaning.

To our knowledge, this research is the first attempt to leverage external structured data about individual users' interests in order to improve disambiguation tasks. In particular, this paper makes the following contributions:

1. We develop a model that matches cross-site user identities in order to represent personal interests in terms of structured web resources.
2. We present a system that incorporates this user-specific information in order to increase the coverage of relevant topics when contextual data is sparse.
3. We compare the performance of our strategies with those of existing disambiguation systems, and we also evaluate our system's ability to correctly identify entity meaning against a ceiling provided by human judges.

The remainder of the paper is organized as follows. Section 2 lays out relations to extant research. Section 3 details our approach and the implemented system, called RESLVE (Resolving Entity Sense by LeVeraging Edits). Section 4 describes the experiment we performed to evaluate the system, compares its performance to alternate methods, and provides discussion of results. Finally, Section 5 offers concluding remarks and directions for future research.

2 Theoretical Motivations & Related Work

2.1 Wikipedia: Augmenting entity representations with structured semantic knowledge

Wikipedia has received considerable attention as a knowledge resource that can facilitate text extraction, recognition, and disambiguation tasks. Prior research has been successful in enhancing the representation of concepts and the relations among them by utilizing Wikipedia as a content inventory and by drawing on its organization schemes.

Bunescu and Pasca (2006) and Cucerzan (2007) were among the first researchers to map named entities to Wikipedia resources corresponding to the same concepts. Other early research successfully used the content, titles, and categories of Wikipedia articles to measure semantic relatedness between words (Strube and Ponzetto, 2006).

Since then, subsequent research has continued to identify additional Wikipedia-based features that can serve as external background knowledge to improve NER and NED. Notable examples include… (*see end of this paper for outline of most relevant recent research that need to summarize and insert here*).

Other research has addressed entity resolution by combining an entity's lexical context with a priori information about a candidate resource's prominence (Fader et al., 2009; Hoffart et al., 2011).

Most recently, research has demonstrated that these ideas to connect named entities to structured resources can be applied effectively on short texts as well (Boston, 2012; Ferragina, 2012).

However, while much research has explored and incorporated various pieces of information from Wikipedia in order to represent entity context and score similarity, there is an extremely valuable ingredient that has yet to be taken advantage of: **author revision history.** We therefore explore personalizing a popularity-based approach that selects the candidate most prominent in a user interest model.

2.2 User Produced Content: Modeling user interest from online contributions

Prior research has found that a user produces text content both on social media and on Wikipedia that is tied to her personal areas of interest and that these interests can be represented using structured semantic data.

Researchers have found that the linguistic features of a user's tweets comprise the main topics of interest to that user (Pennacchiotti and Popescu, 2011). In addition, research has demonstrated that a user's interests are exhibited by the named entities that appear frequently in Tweets and that they can be represented according to the Wikipedia Category page that covers these entities (Michelson and Macskassy, 2010). Abel et al. (2011) had success in extracting entities from a user's tweets and represent them using Linked Open Data.

Meanwhile, other research has shown that Wikipedia editors seek out articles to edit based on topics of personal interest, and most Wikipedia editors have at least one area of concentration in terms of content categories (Wattenberg et al., 2007). Editor revision histories on Wikipedia have also been used to determine a user's topics of expertise (Lieberman et al., 2009).

Further, not only is article editing behavior indicative of topical interest (Cosley et al., 2007) but research has demonstrated that Wikipedia resources such as article pages and category graphs can effectively represent the concepts corresponding to these interests (Syed et al., 2008).

Consequently, we pursue the hypothesis that an individual's topics of interest are expressed equitably across multiple sites and that by connecting that user's identities on social media and Wikipedia, we can exploit the latter's large amount of structured knowledge to augment the sparse contextual information available on the former. We discuss this idea of connecting a user's online identities in the next section.

2.3 Virtual Identity Alignment: Matching cross-site accounts to a single unique user

Research has shown that it is possible to detect and connect profiles belonging to the same individual across multiple online systems. Individual users hold accounts and participate at multiple online communities, and it is possible to uniquely identify users given the presence of matching profile information.

Recent work found that if certain profile attributes such as name, email address, or hometown are available and consistent, then it is possible to identify with high precision accounts on different sites belonging to the same user. (Abel, 2011). Carmagnola and Cena (2009) examine various similar identification properties in order to address cross-system user identification.

Most notably, Perito et al. (2011) demonstrate that it is possible to identify accounts belonging to the same individual with high precision solely based on matching usernames.

(*Also need to look at Szomszor, 2008; Zafarani, 2009; Vosecky et al., 2009; Iofciu et al., 2011; Shen et al., 2012…)*

2.4 Implications

Combining and building on all of these ideas, we thus set out to determine the correct sense of an ambiguous entity detected in a short text on a social media website via the following approach, which is detailed further in the next section:

1. We map the ambiguous entity's candidate senses to their corresponding structured resources in Wikipedia.
2. We identify the Wikipedia account that belongs to the same person who authored the ambiguous text by examining and connecting public information found in account profiles on Wikipedia and the social media service.
3. We determine the revisions made by the user on Wikipedia and collect the corresponding structured resources.
4. We measure the overlap and similarity of this structured information with that associated with each candidate sense from step 1.
5. We select the sense with the greatest overlap as the intended sense.

3 Approach and Implementation

(*this section needs much better explanations written for it as the results come back and the algorithms get finalized..)*

3.1 Representing Authors and Entities with Wikipedia Resources

Our goal is to model the user's domains of interest with respect to Wikipedia resources and its knowledge organization scheme. To do so, we collect a subset of the user's recent edits and the resources those edits affected, ignoring the following kinds of edits that are considered either trivial or irrelevant:

- Revert edits to fix vandalism, typo correction (Cosley et al., 2007).
- Edits on list pages and category pages (Fader et al., 2009) as well as disambiguation pages (Gabrilovich and Markovitch, 2006) as these are not article pages that correspond to entity concepts.
- More of these based on article attributes ignored by Gabrilovich & Markovitch?
    - containing less than 100 non-stop words
    - containing fewer than 5 incoming and outgoing links
    - that describe specific dates
- "Minor" edits

Elizabeth Murnane 10/1/12 1:33 AM
Comment [1]: Define this according to Wikipedia API's definition

Then given this article page, we aim to enrich its feature set with additional related Wikipedia resources. In this paper, we therefore explore variants of the representation that are based on:

- Each article as a Bag of Words
- All direct categories of all articles
- The full category hierarchy graph consisting of the article's direct categories, those categories' categories, and so on.

Our representation for an ambiguous named entity is based on the same ideas. Specifically, each of the entity's candidate meanings is mapped to the Wikipedia article corresponding to that candidate's concept as well as that article's direct parent categories and the full category hierarchy graph.

We now move on to describe the RESLVE system, which implements a variety of approaches based on the various attribute-enriched representations just described.

3.2 Measuring Similarity: The RESLVE System

We refer to our scoring functions for candidate selection and entity resolution as the RESLVE system. The functions are WSD and VSM based and are designed to utilize the various potentially effective attributes of entity mentions and user interest models in order to discover the resulting effects on accuracy and performance.

Bagga & Baldwin (1998) introduced using the Vector Space Model to address NED by representing the context around a named entity with word vectors that could be compared using cosine similarity. We explore 3 VSM based functions:

- VSM_ArticleBOW
  - Builds an inverted index from words to the articles that contain them
- VSM_ParentCat
  - Each article in Wikipedia belongs to one or more categories, which are listed as outgoing "Category:" links on an article's page.
- VSM_CatGraph

We represent a user's edits and the entity's candidates as attribute vectors with TF-IDF based weights. Note that we consider the number of times a user has performed edits on a given article as a contributor to that article's significance, so we incorporate this number into the weighting scheme to give more preference to resources that seem to receive more of the editor's attention. We determine the candidate meaning to select by computing the cosine similarity between the user vector and each candidate vector.

Our WSD approach is based on the Lesk algorithm (Lesk, 1986), which measures the amount of overlap between the dictionary definition of a word with the definitions of the other words in its neighboring context. Specifically, our WSD algorithm selects as the correct candidate for an ambiguous entity the one that has

the highest overlap between that candidate's feature set and the author interest model feature set.

Finally, as Fader et al., 2009 point out, an entity may actually refer to a less prominent candidate meaning. Therefore, we also incorporate context information from the baseline disambiguation function, and we select the baseline disambiguation function's candidate if this lexical context information is above a certain threshold.

4 Experiment and Evaluation

4.1 Data Gathering

Over _____ months we retrieved recent revisions made on Wikipedia and the editors who made them in order to build a dataset of usernames corresponding to _____ unique and active Wikipedia accounts. For each of these accounts, we then downloaded their Wikipedia contribution histories, filtering out invalid revisions according to the method described in section 3.1 until we had obtained the 100 most recent and unique article resources on which the user had made a valid edit. Users for whom at least 100 such resources were not available were removed from the sample, resulting at this point in a dataset of _____ valid usernames.

We then performed cross-site username matching on the social media sites Twitter, _____, and _____. For each site, this involved checking first whether the username existed on the site and if so, whether that user had posted at least 100 pieces of public content on the site. Removing usernames that did not meet these criteria left a dataset of _____ usernames.

The final step in building our dataset of users involved using Mechanical Turk in order to obtain human confirmation that a given username that existed on both Wikipedia and on a social media site actually belonged to the same single individual. "Qualified" Mechanical Turkers were given the URL of the Wikipedia User Page corresponding to that username as well as the URL pointing to the profile of the username on the social media site. They were instructed to compare the Wikipedia profile and the social media profile for agreement in available information such as real name, email address, profile picture, location, or occupation in order to confirm, reject, or deem indeterminable whether the username belonged to the same person. Each username was given to ___ different Turkers, and usernames for which every evaluation was not a positive confirmation that the owner of each account was the same individual were removed from the dataset. This resulted in a total of _____ valid usernames.

Elizabeth Murnane 10/1/12 1:16 AM
Comment [2]: Is that the right term?

Upon obtaining this set of validated usernames, we downloaded the 100 most recent content posted by the user on the social media site for which the

username was confirmed in the previous step. To build a set of all mentions of named entities in a short text, we utilized the Wikipedia Miner Search Service[1], an established named entity recognition toolkit that maps named entity concepts to Wikipedia articles. The service also calculates the probability that a named entity actually refers to a given article based on measures of [..., ..., ...], offering a probability ranked list of candidate articles for an ambiguous entity. Any named entity that the Search Service was unable to map to any Wikipedia article was ignored, as was any named entity that mapped to only a single candidate article (i.e., was unambiguous). Any short text thus not containing at least one ambiguous entity was removed from the dataset. Table ____ shows the resulting numbers of valid short texts and ambiguous named entities obtained from the various social media sites.

Elizabeth Murnane 10/1/12 2:15 AM
Comment [3]: Check their codebase

	# valid short texts	average # of entities detected in a single short text	average # of ambig. entities detected in a single short text	max # of ambig, entities detected in a single short text	min # of detected entities in a single short text
Twitter					
Flickr					

Table ____

At this point, we again turned to Mechanical Turk in order to obtain human judgments of the correct candidate meaning of an ambiguous named entity. We again only utilized qualified Turkers and this time also required native English speakers, as we wanted to ensure that they would be capable of making correct judgments about potentially highly nuanced meaning.

For each ambiguous entity, we asked _____ different Turkers who met our qualifications to read a short text containing an ambiguous named entity and to then choose the correct meaning of that entity from a randomly ordered list of candidate meanings, which we obtained from Wikipedia Miner.

Finally, for each ambiguous named entity, we applied our personalized user-model based disambiguation algorithms described in section ____ to the candidate meanings in order to select a single candidate as the user's intended meaning. The next section details the performance of each of these algorithms and compares the results to baseline and ceiling measurements.

4.2 Results

We use Wikipedia Miner's top ranked candidate as our baseline and the judgments of humans on Mechanical Turk as our ceiling. We evaluate the RESLVE system's disambiguation algorithms according to their accuracy, which is the

[1] http://wikipedia-miner.cms.waikato.ac.nz/services/

fraction of correctly disambiguated named entities out of the total number of ambiguous entities analyzed (i.e., the degree to which the RESLVE system agrees with the evaluations made by human judges).

Algorithm / Dataset	RESLVE-BOW-VSM	RESLVE-BOW-WSD	RESLVE-ParentCat	RESLVE-CatGraph	RESLVE-solo	RESLVE-combo
CUTwitter12						
CUFlickr12						
…						

Table: Comparison of the accuracy of our system's various disambiguation functions

Method / Dataset	Baseline	RESLVE	Ceiling	Improvement
CUTwitter12				
CUFlickr12				
…				
…				

Table xx: Comparison of the accuracy of our system's best performing technique against a [random baseline/wikiminer baseline] and a human judge ceiling

Our actual coverage is x% because as explained in section 4.1, we ignore detected entities that do not map to more than one candidate meaning and so x = number entities not ignored / (ignored entities + not ignored entities)

[Insert here: Scatterplot where x axis is Document Length and y axis is accuracy. Each of 3 algorithms (Baseline, RESLVE, Ceiling) gets a different "dot". Data points should form lines where our accuracy is higher for short texts and then maybe converge with longer text.]

Figure ___: Average accuracy on Twitter dataset with xx users and varying short text length

4.3 Discussion of Findings

Need to do this part once have more results...when is it particularly good? (very short texts?) when does it fail? what kinds of entities? Upon manual inspection of the top and bottom results we realize…..

5 Conclusions and Future Work

More and more, rich records of individuals' online activities and discussions are becoming available on the web and offering new opportunities to determine the semantics of users' text based content based on their past online contributions. In response, we presented a novel solution to the problem of entity resolution in short texts by connecting user identities across communities in order to model the interests of short text authors and apply such models as a source of entity context. Our approach is based on the idea that an individual possesses a core set of interests that is propagated and expressed across multiple online communities in which they participate. We showed that it is possible to improve disambiguation techniques by confining the search space for entity meanings to those topics that a user has previously exhibited such interest in.

A few directions for future work stand out. Currently, we are unable to handle the case when the same entity occurs more than once in same short text?? Because we do not incorporate any positional information about an entity, we would not be able to distinguish in this case?? We would suggest same probability for both and might actually hurt baseline performance?? It would also be desirable to automate the process by which usernames existing on multiple sites are evaluated to determine whether they belong to the same person. A next step could to be to employ the methods developed by (cite) or develop classifiers that utilize the labeled data we now have in order to develop systematic methods for cross-site user identity matching.

Elizabeth Murnane 10/1/12 2:50 AM
Comment [4]: Figure out other limitations

This paper introduced the notion of personalizing named entity disambiguation.

- We presented a novel approach to the challenge of disambiguating named entities in short texts by enriching traditional representations with user-specific information.
- We demonstrated the feasibility of connecting online user identities in order to exploit external structured knowledge that is specific to the author of an ambiguous entity.
- We introduced a novel system that implements the new strategies, and we compared its performance to other state of the art techniques, providing evidence that this approach is effective and holds promise for the future.

References

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. Presented at the UMAP'11: Proceedings of the 19th international conference on User modeling, adaption, and personalization, Springer-Verlag.

Abel, F., Herder, E., Houben, G. J., Henze, N., & Krause, D. (2011). Cross-system user modeling and personalization on the social web. User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems, 22(3), 1–42.

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

Bunescu, R., & Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. Presented at the Proceedings of EACL.

Carmagnola, F., & Cena, F. (2009). User identification for cross-system personalisation. Information Sciences, 179(1-2), 16–32. doi:10.1016/j.ins.2008.08.022

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2006). Using intelligent task routing and contribution review to help communities build artifacts of lasting value. Proceedings of the SIGCHI conference on Human factors in computing systems, 1037–1046.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. Presented at the Proceedings of EMNLP-CoNLL.

Fader, A., Soderland, S., Etzioni, O., Center, T. (2009). Scaling Wikipedia-based named entity disambiguation to arbitrary web text. WikiAI09 Workshop at IJCAI.

Ferragina, P., & Scaiella, U. (2012). Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Software*, *29*(1), 70–75. doi:10.1109/MS.2011.122

Fleischman, M. B., & Hovy, E. (2004). Multi-document person name resolution. Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Reference Resolution Workshop.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. Proceedings of the National Conference on Artificial Intelligence, 21(2), 1301.

Gentile, A. L., Zhang, Z., Xia, L., & Iria, J. (2009). Graph-based semantic relatedness for named entity disambiguation.

Han, X., & Zhao, J. (2009). Named entity disambiguation by leveraging wikipedia semantic knowledge. Proceedings of the 18th ACM conference on Information and knowledge management, 215–224. doi:10.1145/1645953.1645983

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., et al. (2011). Robust disambiguation of named entities in text. Presented at the EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Kazama, J., & Torisawa, K. (2007). Exploiting Wikipedia as External Knowledge for Named Entity Recognition. EMNLP-CoNLL, 698–707.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation, 24–26.

Lieberman, M. D., & Lin, J. (2009). You are where you edit: Locating Wikipedia contributors through edit histories. Proceedings of ICWSM, 9.

Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. Presented at the KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Request Permissions. doi:10.1145/1557019.1557073

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. Presented at the AND '10: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, ACM Request Permissions. doi:10.1145/1871840.1871852

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. Presented at the Proceedings of NAACL HLT.

Mihalcea, R., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. CIKM, 233–242. doi:10.1145/1321440.1321475

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

Nguyen, H., & Cao, T. (2010). Exploring wikipedia and text features for named entity disambiguation. Intelligent Information and Database Systems, 11–20.

Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to twitter user classification. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM).

Perito, D., Castelluccia, C., Kaafar, M. A., & Manils, P. (2011). How unique and traceable are usernames? Presented at the PETS'11: Proceedings of the 11th international conference on Privacy enhancing technologies, Springer-Verlag.

Pilz, A. (2010). Entity disambiguation using link based relations extracted from wikipedia. Automated Knowledge Base Construction, 37.

Shen, W., Wang, J., Luo, P., & Wang, M. (2012). LINDEN: linking named entities with knowledge base via semantic knowledge. Presented at the WWW '12: Proceedings of the 21st international conference on World Wide Web, ACM. doi:10.1145/2187836.2187898

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. AAAI, 1419–1424.

Syed, Z., Finin, T., & Joshi, A. (2008). Wikipedia as an ontology for describing documents. Proceedings of the Second International Conference on Weblogs and Social Media, 136–144.

Wattenberg, M., Viégas, F., & Hollenbach, K. (2007). Visualizing activity on wikipedia with chromograms. Human-Computer Interaction–INTERACT 2007, 272–287.

Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, 197–205.

Need to summarize and insert this in section 2.1:

- Bunescu & Pasca, 2006
  - Treat NED as a ranking problem
  - Define a measure of similarity between the words in the context of an entity and the words in the Wikipedia pages corresponding to its candidate meanings.
  - First attempt basic TF-IDF cosine-sim with article text, then augment article's term vector with additional words from other articles in the same category
  - Also use Wikipedia articles, categories, hyperlinks, redirect pages, disambiguation pages to determine relation between named entities.
- Mihalcea, 2007

    - Generalize Bunescu & Pasca beyond named entities, utilizing article source and target links for keyword extraction and disambiguation.
- Strube & Ponzetto, 2006
    - Successfully utilize Wikipedia articles' content, titles, and categories to measure semantic relatedness between words.
- Gabrilovich & Markovitch, 2006
    - Use Wikipedia for document classification
    - Build inverted index from words to articles containing those words
    - Formulate relatedness score between documents by mapping each document to a Wikipedia article and using the inverted index to enrich each document's representation.
    - Compute cosine similarity between document vectors
    - Paper demonstrates the feasibility of augmenting machine learning classification techniques with external knowledge relevant to the text being processed.
- Cucerzan, 2007
    - Uses a Vector Space Model representation
    - Employs an entity's surface forms, category tags, and contextual information like co-occurring terms as disambiguation clues.
    - Disambiguation strategy maximizes overlap between document vector of ambiguous entity and each candidate vector
- Milne, 2008
    - Following from Mihalcea, defines relatedness in terms of overlap of incoming links in Wikipedia articles
- Fleischman & Hovy, 2004; Bagga & Baldwin, 1998
    - Similarity using bag of words model that represents entity's context in weighted term vectors
- Nguyen & Cao, 2008
    - Explore and incorporate a number of features from Wikipedia in order to represent entity context for similarity scoring: article title, redirect pages' titles, category titles, incoming & outgoing link text
- Gentile et al., 2009
    - Builds features based on various pieces of information related to a concept's corresponding article: words in the article's title, categories, outgoing links, and the most frequent words on the article's page.
- Han, 2009
    - Represents named entities as concept vectors and computes the similarity between them
    - Augments keyphrase representations with hyperlink information
- Kulkarni et al., 2009
    - Wikipedia as an entity catalog
    - "Coherence prior"
- Popularity-based
    - Fader et al., 2009

        - Use a priori prominence information in combination with contextual information to disambiguate
        - Entity may actually refer to a less prominent candidate, so select candidate with highest prior probability unless the contextual information is above a threshold
    - Hoffart et al., 2011
        - Use popularity-based and graph-based models to measure similarity between mentions and entities.
        - Define prominence/popularity in terms of Wikipedia anchor texts and incoming link frequencies
- Pilz, 2010
    - Uses an SVM approach with features such as article text, category, and links to other categories
- Kazama & Torisawa, 2007
    - Demonstrate that category labels extracted from Wikipedia articles can be used to improve accuracy of NER
- Shen et al., 2012
    - Use entity pages, redirect pages, disambiguation pages, hyperlinks to link named entity mentions with a knowledge base that unifies Wikipedia and Wordnet.
    - Use taxonomy of knowledge base to measure semantic similarity between knowledge base concepts and candidate entities