Norwegian University of Science and Technology
Technical Report IDI-TR-1/2009

Extracting Named Entities and Synonyms from Wikipedia for Use in News Search

Christian Bøhn and Kjetil Nørvåg*
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

ISSN: 1503-416X

*Email of contact author: Kjetil.Norvag@idi.ntnu.no

Abstract

In many search domains, both contents and searches are frequently tied to *named entities* such as a person, a company or similar. An example of such a domain is a news archive. One challenge from an information retrieval point of view is that a single entity can have more than one way of referring to it. In this paper we describe how to use Wikipedia contents to automatically generate a dictionary of named entities and synonyms that are all referring to the same entity. This dictionary can subsequently be used to improve search quality, for example using query expansion. Through an experimental evaluation we show that with our approach, we can find named entities and their synonyms with a high degree of accuracy.

1 Introduction

In many search domains, both contents and searches are frequently tied to *named entities* such as a person, a company or similar. An example of such a domain is a news archive. One challenge from an information retrieval point of view is that a single entity can have more than one way of referring to it. In some bases, this can be the result of sometimes using the abbreviation versus not using it (e.g., *United Nations* and *UN*), other times it will be because of different ways of referring to a person (e.g., *Barack Obama* versus *President of the United States*). The use of the different search terms will give very different search results and ranking of search results, since the search engine treats the queries as if the user was interested in a particular spelling, even if they from a user's point of view might refer to the same entity and should be considered equivalent.

In order to improve search quality in such domains, it will be useful to 1) recognize such named entities (NEs) in the text, and 2) determine possible synonyms for each NE. This knowledge can subsequently be used to increase search quality by the use of query expansion, where a search for a NE can also give the results of synonyms of the entity. In order to handle emerging names, freshness of the NE/synonym dictionary is important, i.e., as soon as new NEs emerge (for example previously unknown persons) they should be included in the dictionary.

Wikipedia has most likely become the largest freely available collection of knowledge and as of January 2009 it contains more than 2.7 million articles in the English version [14]. In this paper we will explore the idea of using Wikipedia contents to automatically generate a dictionary of NEs and synonyms that are all referring to the same entity. With such a dictionary in hand we can then handle entities in a way so that the spelling of the entities becomes less important, making it possible for the search engine to return potentially interesting news articles mentioning the entity, but with a different synonym.

In the paper we describe an approach for Wikipedia-based NE recognition that significantly improves performance of previous approaches, and describe an approach for determining synonyms among the NEs. An experimental evaluation confirms that Wikipedia is well suited as a source of NEs and synonyms aided by its semi-structuredness that can help in recognizing entities and related synonyms, and that entities can be found with a very high precision. The proposed method also classifies the NEs as *people*, *organizations*, and *companies*. The categories can for example be used by users to filter search results according to what they are searching for.

Thus our main contributions of this paper are 1) an approach for improved NE recognition that also implicitly categorizes the found entities, 2) discovery of synonyms of the NEs, 3) overview on how the synonyms have been used in our system to improve search quality, and 4) a study of quality of NE extraction and synonym discovery. The combined advantages of the approach include language-independency, unsupervised, not rule-based, and no need for manually annotated training data. Since it is simply based on processing Wikipedia it means that it can provide freshness: new named entities can be included in the directory as soon as they appear in Wikipedia.

The organization of the rest of this paper is as follows. In Section 2 we give an overview of related work. In Section 3 we describe the contents of Wikipedia as well as a generic NE recognition algorithm. In Section 4 we describe our improved NE recognition algorithm. In Section 5 we describe how to perform synonym extraction. In Section 6 we describe how to use query expansion when performing searches. In Section 7 we describe experiments and

results. Finally, in Section 8, we conclude the paper and outline issues for further work.

2 Related Work

During the recent years several attempts have been made in using the semistructured contents of Wikipedia for information retrieval purposes. The ones most relevant to our work are [4, 12, 15].

In [15] Zesch et al. evaluate the usefulness of Wikipedia as a lexical semantic resource, and compares it to more traditional resources, such as dictionaries, thesauri, semantic wordnets, etc. In [4] Bunescu and Paşca study how to use Wikipedia for detecting and disambiguating NEs in open domain text. Their motivation is to improve search quality by being able to recognize entities in the indexed text, and disambiguate between multiple entities that share the same proper name by making use of the context given by the text. Then during searches they want to group results according to sense rather than as a flat, sense-mixed list. That would give the users access to a wider range of results as today's search engines may easily favor the most common sense of an entity, making it difficult to get a good overview of the available information for a lesser known entity. In order to recognize NEs, they use a simple, three-steps heuristics which we also build upon in our paper (Section 3.2). Next, they use the redirect pages to find alternative names for the entities, and disambiguation pages are used to identify different entities that all share the same proper name. Similar ideas have also been used by Cucerzan [5]. In [12] Schenkel et al. present their system YAWN, which converts Wikipedia into semantically annotated articles. Their motivation is to open up for a more advanced query syntax, making it possible to use semantically rich structural queries that are very precise in what they are looking for like //person[about(//work,physics(and about(//born,Germany)] to query Wikipedia.

Knowledge from Wikipedia can also be used in order to improve the quality of traditional NE approach. Kazama and Torisawa [8] describes how to extract categories from the first sentence in a Wikipedia article and using these categories to improve NE recognition.

Another useful public available source of semantic information is WordNet, a large lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, or *synsets*, expressing distinct concepts[6]. Synsets are linked, based on conceptual-semantic and lexical relations. In [11], Magnini et al. describes a method where WordNet is used to create a collection of NEs, grouped by categories such as *person*, *location*, *organization*, etc. Their method is based around capturing external and internal evidence, where internal evidence are the words in the text that are considered to be an entity and the external evidence are the surrounding sentence. Toral et al. used this as the basis for their NE extraction in [13].

Although NE recognition is nothing new, traditionally the focus has been on recognizing NEs embedded in text. Most approaches are based on rules [3], decision trees [10], hidden Markov models [2], and maximum entropy [7]. However these methods do not take into account the additional semantic information available due to the Wikipedia structure and might also be too time-consuming when the aim is to have a dynamic dictionary which is continuously updates based on the evolving Wikipedia.

3 Preliminaries

In this section we will describe more detail the contents of Wikipedia as well as the generic NE recognition algorithm proposed by Bunescu and Paşca [4].

3.1 Wikipedia

There are four Wikipedia features that are in particular attractive as a mining source when building a large collection of NEs: internal links, redirects, disambiguations, and categories. In the following sections we will briefly describe these features.

Internal Links: Internal links are used to link words in one article with another article, thereby making it very easy for the users to find more information about a specific keyword mentioned in the article text.

Redirects: Redirects are almost similar to links, except that they can not include an alternative text. We intend to use them as another source of synonyms or alternative spellings of entities, as was done in [15]. A difference between redirects and links are that the links pointing to different articles can share the same display text, but a redirect can only redirect to a specific article. This makes the redirects less ambiguous. An example of a redirect is to redirect *Shortest path* to *Shortest path problem*.

Disambiguations: Disambiguation[1] pages are used by Wikipedia to resolve conflicts between terms having multiple senses by either listing all the senses for which articles exist, or treat the dominant sense as the primary article, and then presenting a small link to less popular senses. An example of an ambiguous term is *Mercury* which can refer to both the element and the planet as all Wikipedia article titles start with a capital letter.

Categories: Categorization is used to group one or more articles together, and every article should be a member of at least one category. However, this is only encouraged, not required. The categories that a page is a member of are always shown at the bottom of the article, and can help the users in finding other articles related to the domain. The categorization system is flexible as it is not limited to a tree structure, instead it is a direct cyclic graph. While avoiding cycles is encouraged, it is not enforced by the software and therefore some cycles exist. This may make it difficult to determine which category is the parent category and which one is a sub-category.

3.2 Generic Named Entity Recognition

When using Wikipedia titles for NE recognition, a first naive attempt might be to use capitalization of words to find entities. However, this approach will not work because all article titles have their first letter capitalized even if they are nouns rather than proper nouns. A more sophisticated approach is described by Bunescu and Paşca in [4], and is based on the following heuristics:

[1] Note that the meaning of the term *disambiguation* in Wikipedia context is slightly different from how it is used in computational linguistics.

- If multi word title and every word is capitalized, except prepositions, determiners, conjunctions, relative pronouns or negations, consider it an entity.
- If the title is a single word, with multiple capital letters, consider it an entity.
- If at least 75% of the occurrences of the title in the article text itself are capitalized, consider it an entity.

A fixed value of 75% is not necessarily robust, so we instead use a more flexible threshold approach where the title is considered an entity if the fraction of capitalized occurrences of the title in the article text is larger than α. We will later in this paper study experimentally how different values of α affect the NE recognition.

4 Improving Named-Entity Recognition

Although the generic approach for NE recognition described above has relatively good performance despite its simplicity, improvements can be made. In this section, we describe an alternative to the capitalization requirement, which utilizes Wikipedia categories that are mainly made up of entities, to recognize NEs.

As mentioned above, the Wikipedia categories form a directed cyclic graph, which makes it more difficult to find nodes in the category graph that designates that all sub-categories are people, organizations, or companies. Since it does not follow a tree structure, we risk running into cycles, which could turn the remaining of a graph into a sub-category of a chosen parent node. In order to avoid this problem, we instead use the fact that category names often follow certain patterns when multiple categories are related. For instance, there are multiple category groups that follow a *Companies based in xxx* pattern, where *xxx* is a geographical location. We believe that this can possibly be very useful for gathering a large collection of entities related to a few groups. Also, this collection of entities will be useful in evaluating the recall of the entity recognition algorithm described above.

Since we intended to use the extracted entity dictionary in a news context, we selected three categories of entities we consider highly relevant for our intended application areas: 1) people, 2) organizations, and 3) companies.

The first entity category is easy to find entities for, as there is a category named "'Living people."' This category exists in Wikipedia mainly because living people may suffer harm if wrongful information is attributed to them, and therefore these pages must be watched more carefully than other pages. This makes it a very useful category to us as it should cover most people who are news relevant.

The second and third entity categories are more difficult to extract as there are no superior category for either of them which are used to indicate that all children are either organizations or companies. In order to solve this problem, we use pattern matching to identify categories holding entries that would fit under the respective NE categories. Using simple wildcards we found category patterns that matched categories that are made up of entities, as shown in Table 1 where the patterns we used are listed.

It should be noted that a side-effect of our category-based NE recognition approach is a set of classified NEs. This can also be utilized in order to increase search quality.

Entity Category	Pattern
Companies	"'Companies headquartered in *"'
Companies	"'Companies established in *"'
Companies	"'Companies based in *"'
Companies	"'Companies listed on *"'
Companies	"'* companies of *"'
Companies	"'* companies"'
Organizations	"'* organizations"'
Organizations	"'Organizations based in *"'
Organizations	"'Organizations established in *"'
People	"'Living people"'

Table 1: Patterns used for category matching.

5 Synonym Extraction

After a set of NEs have been identified, we want to find their synonyms. We intend to use the internal links, redirects and disambiguation pages for this, and we can easily extract all of these after we have the NEs. This will give us a list of captions, all used on links to a particular entity. The list can contain two types of "noise": 1) it is likely to contain a various amount of "junk synonyms", i.e., synonyms that are not really synonyms, but instead the result of people vandalizing articles, and 2) it can also contain link captions where a noun has been appended to the proper noun which it is linking to, e.g., *Bush administration's* is linking to the article about *President Bush*, yet it is not a good synonym.

To filter out the noise we considered two options:

- Weighting each link caption based on the number of links using the same caption. Then we can filter out the less popular ones, which are less likely to be good synonyms since they are used infrequently.
- Apply the same algorithm used for classification of link captions, where we use versions of the link captions that are capitalized in different ways as an alternative since we have no article text.

In the synonym extraction step we want to extract all the possible synonyms for all the NEs we had identified earlier. We collected all the links and redirects with destination and caption. Since we are not interested in the source article, we accumulated all links pointing to the same title, using the same caption. The synonyms listed in Table 2 are an example of what we found through the synonym extraction. The synonyms listed here and their frequencies are real, but the selection of synonyms was done manually in this case.

Unfortunately the links do not provide us with a perfect set of synonyms as the link captions in some cases are very contextually dependent. What this means is that we found link captions pointing to NEs were the link was made up of a pronoun or other terms than proper nouns. In some cases the entity name used in the link caption is not even the same entity that the link is pointing to, instead they are only related in some way. To deal with some of the noise we apply filtering as follows:

- Given the set S of potential synonyms for an entity, for each $s_i \in S$:

Main Name	Synonym	Frequency
George W. Bush	George W. Bush	7166
	Bush	453
	President Bush	392
	George Bush	129
	President George W. Bush	65
	G.W. Bush	62
	George W. Bush	32
United Nations	United Nations	9943
	UN	816
	U.N.	88

Table 2: Example of a synonym set.

- Remove any suffix enclosed in parentheses and apply a light stemming stripping it of any possessive form
- Classify the synonym as good or bad synonym as described below, remove s_i from S if it turns out to be bad
- Given $freq(s_q)$ as the frequency of a synonym s_q and |S| as number of items in S, remove s_i if $freq(s_i) < \sum_{k=1}^{|S|} freq(s_k) * \beta$
  (In our experiments we have used $\beta = 0.01$)

When trying to classify the synonym as a good or bad synonym we use a similar algorithm as the one described in 3.2, except we do not have an article text with occurrences we can use, therefore we ignore that rule. Since we then lose the rule which was used to handle single-word names, we lower the limit of the minimum capitalized words required to one. We also use the frequency of a potential synonym to weight its importance and remove the ones that fall below a given threshold.

6 Employing Synonyms in Search

One of the motivations for automatically building a dictionary of NEs and their synonyms is to use it in order to improve robustness in searches, by being able to improve the recall when entities are referred to using different names in the query and the documents, e.g., *United Nations* and *UN*.

How to utilize the dictionary of NEs depends on whether we can have our framework inside the search engine (internal) or only as a frontend (external). The former is feasible in a enterprise/institutional search domain, while only the latter is feasible when using search engines like Google and Yahoo.

External. In the case of accessing external search engines, the problem is that we do not have access to the the original news articles and the ability to normalize [9] them before indexing.

A first näive approach in using our approach as frontend to an external search engine is to use query expansion by expanding the query to include multiple synonyms. However, as was also witnessed in our experiments, this will not give very good results. Actually, few of

the original highest-ranked results from the non-expanded version of the query (which should be expected to also appear in the query-expanded result) will be in the result set. The reason for the problem of using this approach relates to how the vector space model works: search results will only be ranked high when the results contain multiple synonyms in the same text. Although in general this problem can be solved by the use of *Generalized Vector Space Models (GVSM),* when employing external systems we this is not an option (that GVSM are not used in search engines were also obvious from our experiments).

A second and better approach, is to submit individual queries and present a ranked version of the total/merged result set as result to the query. However, this approach also has potential problems. First, for news search engines, there will be a number of duplicate articles, i.e., articles that are not only reporting on the same topic but more or less are the same article. Based on observations from our experiments, it appears that news search engines would arbitrarily remove all but one of the duplicates from the result set during query time. What this meant was that the removed duplicates would change depending on the query used, making it more difficult to do automatic filtering of the merged result set. However, this problem can mostly be solved. A second and more serious problem is how to rank the final result set. For a start, simply merging the individual result sets, is a possibility. However, this technique can be improved by more sophisticated ranking methods.

Internal. In the internal case, an additional approach that can be applied is to perform entity normalization before indexing. That is, all occurrences of an entity are translated into their main entity reference if it can can be determine which entity the document is about, or a list of the unique names of multiple entities if there are no unambiguous references in the text. This has similarities to GVSM.

Synonym selection. A problem with the query expansion is that the popular entities have a very large amount of synonyms with very small variations. For example, the entity with the most synonyms had as many as 153 different synonyms (cf. Table 6). If the queries are expanded with all the synonyms of the entities specified, the result would be queries so large that they would most likely result in a serious performance hit. This will be unacceptable in a real world usage, in particular if external resources are used. The solution is to limit the expanded query to the top k synonyms, where a suitable value for k can be around 5-10. Using our approach, the synonyms to select are the ones with the most inbound links using the synonyms as link captions.

7 Evaluation

In this section we describe the experimental setting and evaluation of our proposed ideas. The goal of the experiments is to study the quality of entity recognition and synonym detection using the Wikipedia-based approaches described above.

7.1 Evaluation Environment

In order to evaluate the ideas we implemented a system for extracting NEs and synonyms from Wikipedia (the results in this paper are based on the January 2008 dump containing

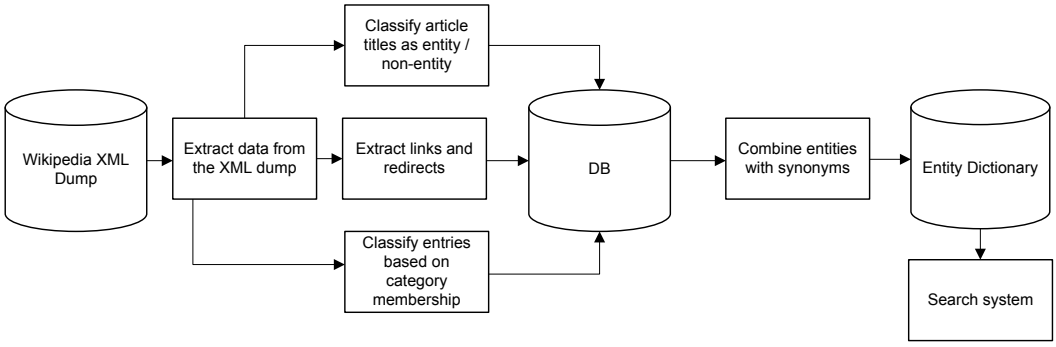

Figure 1: System overview.

the latest version of each article available at Wikipedia's download site[2]). The system also provides a search frontend that employs the NEs and synonyms for increasing quality of searches towards existing search engines. An overview of the system design is shown in Fig. 1.

The metrics used in the evaluation are *precision* and *recall* [1]. The reference set is the set of items that would be generated from the input set if the operation performed on the input set was perfect, so that precision is therefore the fraction of relevant items in the result set, while recall is the fraction of relevant items that were included in the result set. We also employ the F-Measure which combines precision and recall into a single performance measure, i.e., $F = \frac{2*precision*recall}{precision+recall}$.

The paper has a two-fold focus. The first part is automatic generation of a NE dictionary, and the second is using the dictionary to better handle the occurrences of different synonyms. In the first part of the evaluation, where the NEs extracted from Wikipedia are to be evaluated, we extracted smaller subset for in precision/recall calculations. These subsets were randomly chosen and then manually classified. See the appendix for the evaluation data.

7.2 Named Entity Recognition Results

In this section we will present the results of the evaluation of the NEs. First we present the results from the global NE recognition, followed by the results from the three categories we extracted entities from, and last we use the category-based entities to evaluate the algorithm used in the global entity-extraction.

7.2.1 Generic Recognition.

Fig. 2 shows the precision, recall and F-Measure for different values of α (i.e., threshold in the generic NE recognition described in Section 3.2) are shown. Here recall is the percentage of the entries that were recognized as entities, while the precision is the percentage of the entries correctly classified as NEs. The test data we used for this was a random subset of the Wikipedia entries which was manually classified as entity/non-entity and can be found in appendix A.

As the recall drops fairly evenly while the precision improves similarly for different values of α, it is difficult to see what the optimal value of α is. Fig. 2 shows the F-Measure for the

[2] `http://download.wikipedia.com`

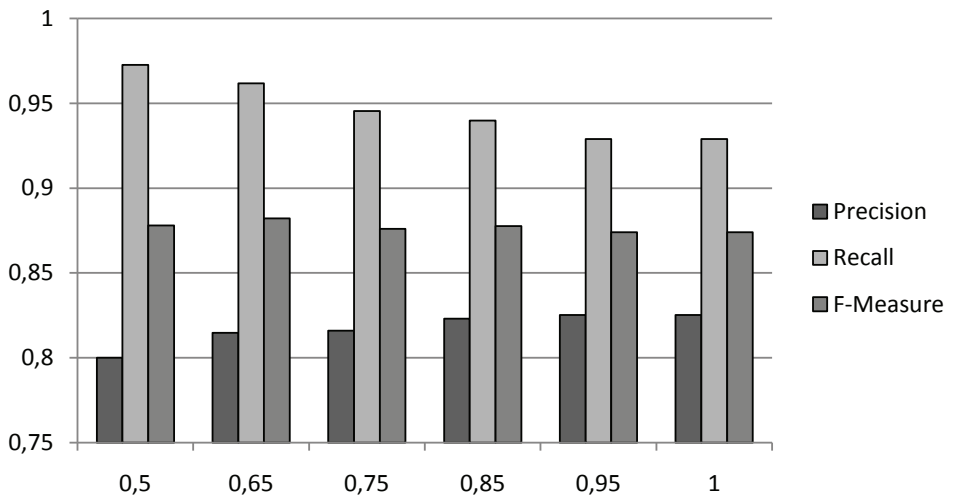

Figure 2: Precision, recall, and F-Measure of the recognized entities for different values of α.

Category	Pattern	Entities
Companies	"'Companies headquartered in *"'	204
	"'Companies established in *"'	7518
	"'Companies based in *"'	8555
	"'Companies listed on *"'	1365
	"'* companies of *"'	15728
	"'* companies"'	10955
Organizations	"'* organizations"'	12661
	"'Organizations based in *"'	1640
	"'Organizations established in *"'	1

Table 3: Number of entities matching each of the patterns.

different thresholds, and shows that based on this combined measure, $\alpha = 0.65$ is the one giving the best results.

7.2.2 NEs from Categories.

The second approach we used to generate lists of entities was based on the use of string patterns to recognize the categories used for different kinds of entities. Table 3 shows a breakdown of how entries matching the different patterns were divided. As one entry can be a member of multiple categories, the total number of entities per category is less than the sum of the entries matched by each pattern, and the number of unique entities per category can be seen in Table 4.

We selected a a random subset of 585 entities that match any of the patterns (the list of entities can be found in Appendix B), and then calculated the precision by manually classifying this subset. From this we found a very small list of entries that were not NEs. These are shown in Table 5. As can be deduced from the names, most of these are in reality entries that list multiple entities or general terms, except for *Albert and David Maysles* which

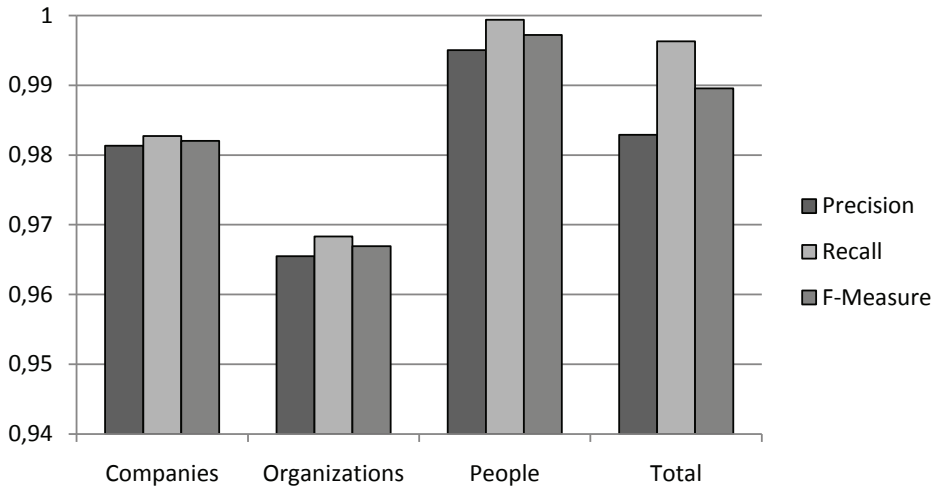

Figure 3: Precision, recall and F-Measure of the categorized entities.

Category	Unique Entities
Companies	27188
Organizations	11988
People	228071

Table 4: Number of unique entities per category.

Category	Non-Entity
Companies	China-based financial stocks in Hong Kong Dynamic packaging List of assets owned by Time Warner List of national and international moving associations Norwegian types of company
Organizations	Charity badge Death squad List of Aikido organizations List of fictional companies
People	Albert and David Maysles

Table 5: Non-entities tagged with entity categories.

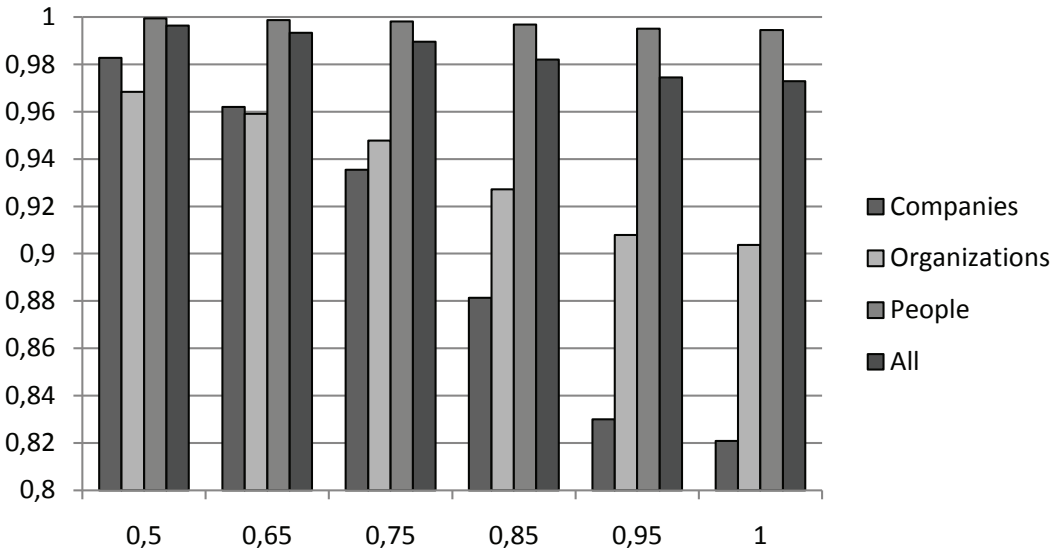

Figure 4: Recall of the NE classification algorithm when used on the categories, for different values of α.

we consider a misclassification still since it is an entry about two different entities that are related, but not a single entity.

In Fig. 3 the precision, recall and F-Measure of the different categories are shown. The recall of the general entity classification algorithm is evaluated using the three categories of entities extracted using the category patterns as test data. Overall the average recall is high since the *people* category is considerably larger than the other two. We have also studied the impact of the α parameter (see Fig. 4), and as can be seen the recall of *companies* and *organizations* varied significantly with the α threshold. The reason is that small uncapitalized words are more common in these entities.

7.2.3 Observations.

Both approaches for extracting NEs from Wikipedia entries have advantages and disadvantages. The first one is a generic method in the sense that it is able to recognize entities from all of Wikipedia. It is based on the fact that proper nouns are capitalized, and NEs are proper nouns. There is one problem, and that is that all Wikipedia entries have the first character in their title capitalized by convention, which means it is not useful to look at the first character to recognize proper nouns. If it was not for that, it would have been considerably easier to recognize NEs with a high precision. Instead we had to rely on a set of heuristics. As seen in Fig. 2 we are able to obtain a precision of 80% and higher with a recall around 95% using these heuristics.

Using the category-based approach yielded a considerably improved precision over the first method, in addition to giving us the entities grouped by categories. The categories selected were categories that are highly related to news search engines or news archives, and the smaller list of entities generated through this method may actually be an advantage. A problem with generating too many entities is that only a fraction of them are actually news relevant and the irrelevant ones may become noise as they match the wrong person. That

Category	# of Entities	Average # of synonyms	Max # of synonyms
Companies	25284	3.2	103
Organizations	11122	2.7	69
People	221207	1.9	153
All	257613	2.1	153

Table 6: Statistics from the synonym extraction.

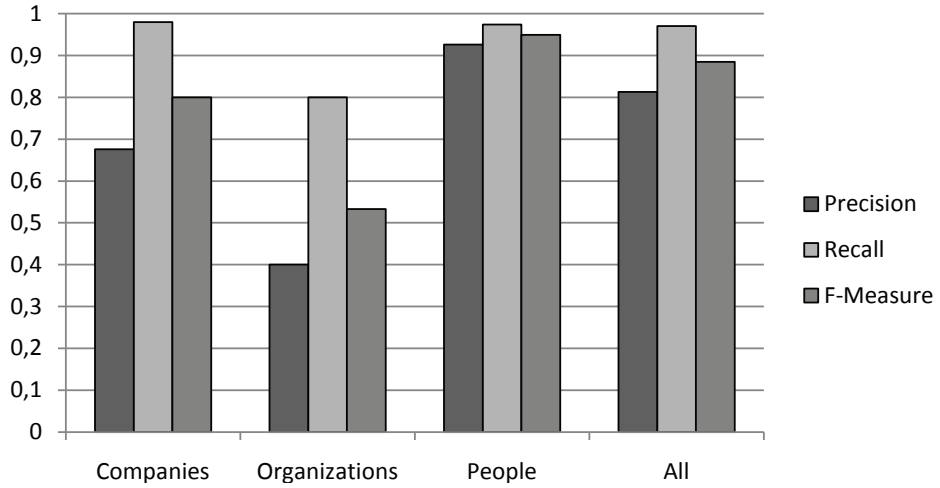

Figure 5: Precision, recall and F-Measure for the synonyms.

is why we selected only a few news-related categories. From what we have seen, it would be fairly easy to use this method to generate a collection of geographical entities, including which entities that are part of another entity simply by looking at the entry's categories and title. In the case of geographical entities, the entry titles often follow a pattern where the things like county, state, or country follow the entity name separated by comma.

7.3 Synonyms

The synonym extraction was based around the categorized entities and the average number of synonyms found per category is shown in Table 6. As we can see the number of synonyms found was in average lower among people than the other categories. We believe this is because of the large amount of people entries in Wikipedia that are very short on content as they are less popular entries, and are therefore having very few links pointing to them. Also, for companies and organizations, the use of abbreviations is more common, resulting in more synonyms on average.

We classified a random subset of the potential synonyms and used this to calculate precision/recall of the link labels and redirects classified as synonyms. As shown in Fig. 5, the precision/recall of people was considerably higher than for companies and organizations. Especially for organizations, the subset used for the evaluations contained very few organizations, which may have affected the precision/recall calculation of this category.

7.3.1 Observations.

Finding synonyms was an important part in the creation of the entity dictionary, and using the entities found earlier we considered all links and redirects to any of them as potential synonyms. What was seen was that the popular entities usually had a very large list of potential entities, often made up of various spelling variations and different uses of abbreviations or titles. One reason for the very large amount of synonyms with very tiny differences is that while the pages for popular entities are of high quality, the same may not hold true for the entries linking to them which results in a lower quality of the link captions coming from these entries. A possible approach to this would be to try to determine the quality of the entries the links are coming from and use that to weight the synonyms.

The results in Fig. 5 indicate a very high precision for the synonyms found for people, but for companies and organizations this is considerably lower. One reason is that companies in some cases have subsidiaries which did not have separate pages, but instead they were only given a short description on the parent company's page. We did not consider this to be the same entity, and therefore filtering the of company synonyms is more difficult than people synonyms. Another explanation is that the people category was very large compared to the other categories, including many short stub articles, and because of this they had fewer average synonyms.

The average number of synonyms listed in Table 6 would have been considerably higher if we had only looked at popular entities. This is to be expected as Wikipedia has more than 200000 people entities, where the majority are not commonly known. These lesser known entities are likely to have very few synonyms.

8 Conclusion

In this paper we have described approaches for using Wikipedia to automatically build a dictionary of NEs and their synonyms. The intended usage of this dictionary is in search by helping the users find articles about the entity independent of which entity name is used in the article.

The evaluation shows that Wikipedia is well suited as a data source for NE mining. We were able to extract a large amount of entities with a high precision, and the synonyms found were mostly relevant, but in some cases, the number of synonyms were very high. This resulted in lots of synonyms that were correct, but would rarely be used in a search query as they were very context specific.

Future work includes using additional Wikipedia structures and contents for improved NE recognition and categorization. One such structure is the template system, where articles can include a template while passing along a set of variables that are used by the template. We also plan to continue to improve the application of the NEs and synonyms in our search frontend, in particular by improving the ranking after synonym-based query expansion has been applied.

Acknowledgments

The authors would like to thank Jon Atle Gulla for helpful feedback in the initial phase of this work, and George Tsatsaronis and Robert Neumayer for valuable help in improving the paper.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] D. M. Bikel, S. Miller, R. M. Schwartz, and R. M. Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'1997*, 1997.

[3] A. Borthwick. *Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, 1999.

[4] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL'2006*, 2006.

[5] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[6] C. Fellbaum. *WordNet – an electronic lexical database*. MIT Press, 1998.

[7] E. Jaynes. Information theory and statistical mechanics. *Physics Reviews*, 106(4):620–630, 1957.

[8] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[9] M. A. Khalid, V. Jijkoun, and M. de Rijke. The impact of named entity normalization on information retrieval for question answering. In *Proceedings of ECIR'2008*, 2008.

[10] D. M. Magerman. *Natural language parsing as statistical pattern recognition*. PhD thesis, Stanford University, 1994.

[11] B. Magnini, M. Negri, R. Prevete, and H. Tanev. A WordNet-based approach to named entities recognition. In *Proceedings of COLING-02 on SEMANET*, 2002.

[12] R. Schenkel, F. M. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *Proceedings of BTW'2007*, 2007.

[13] A. Toral and R. Muñoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Woirkshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[14] Wikipedia, `http://www.wikipedia.org/`.

[15] T. Zesch, I. Gurevych, and M. Mühlhäuser. Analyzing and accessing Wikipedia as a lexical semantic resource. In *Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology*, 2007.

A Classification of Entity/Non-Entity Subset

A.1 List of Entities

Electoral district of South-West Coast
Milo Keynes
Ralph Taeger
International Research on Working Children
Luca Cigarini
"-And He Built a Crooked House-"
LÃL' RÃşisÃ■n (P51)
Britton Johnsen
The Lost Battalion
Royal College Port-Louis (Mauritius)
Rhys Evans
Robert Latta (White House intruder)
Juan Downey
First Restoration
Argentina national rugby union team
Lindsey Wallace
Meezen
Stored Waste Examination Pilot Plant
William Byrd
Attila JÃşzsef
Comte Desbassayns de Richemont
Niemand hÃűrt dich
Tess Bateman
Earshot (Buffy episode)
Simon de Vlieger
Failsworth West
Harry van der Meer
Duane Bobick
Eddy Ko
Hunter Johnson (disambiguation)
Hartvig Svendsen
Emma Roberts
Victoria (New Brunswick electoral district)
Brighton Robins
Dance Got Sick!
Bhawal
Night Skies (film)
Amatole District Municipality
Cadishead
The Killers (short story)
Manuel GutiÃl'rrez NÃąjera
RevÃl's/Yo Soy
Prince Umberto of Bulgaria
Ray Cunningham
Mlynica
Lee Seung-Ho
Elbit Systems
Makyla Smith
War of Genesis
Drumoak
Rupert Holmes
Via dei Fori Imperiali
Arcadia Publishing
Rubens Farias Jr.
Staten Island University Hospital North Campus
Gladius DB
Joseph Alfred Lamy
Zazu
Jacopo Bertucci
West Seneca East Senior High School
Junius Hillyer
Ray Lawson
Pueblo del Arroyo
Philip A. Kent
Carlisle Upperby TMD
Venus (Frankie Avalon song)
Angela Summers
Edward II
Jackie Wright
Castle Ashby
Three Towns
Di Air
Murder on the Nile/Hidden Horizon
My Antonia (film)
Fire in the Abyss
Florida Atlantic Owls baseball
Pierre Ducasse (footballer)
Maria Luisa of OrlÃl'ans
Cleethorpes Pier
Audubon Avenue (Manhattan)
Grand Pass (Washington)
Isabelle Breitman
Buckman Tavern
Playwutchyalike: The Best of Digital Underground

Ken Caillat
Samir El Moussaoui
Chobe National Park
Yusuf Hamied
Timothy Chambers
Onyx 2 On The Bay
James White (General)
UmanitÃă Nova
Fish Leong
W. Allen Wallis
Stewart Reburn
John Wilton
Bernard Herrmann
Barbro Martinsson
Esko Rekomaa
From This Moment On (Cole Porter song)
Trouble at the Henhouse
NetJets
VFinity
Florence Marina State Park
SÃijmeyra Kaya
JMax
Sheriff (band)
Frank Mossfield
Hyderabad District (Pakistan)
Pure Frosting
Imperial College Boat Club
Albedo (Xenosaga)
United States Secretary of Transportation
Andrew Horning
Leszek Dunecki
Lao Bao town
Andrew McGarry
Ballymeanoch
Angkor International Airport
County Route 506 (New Jersey)
Ben Moon
Charles Fickert
A. H. J. Prins
Celestial Season
Martin Evans
Corippo
Alvega
Jacqui Abbott
Emil Molt
Konstantin Mirchev
JosÃl'e Chouinard
Buckeye Municipal Airport
Invisible Ones
Joey Eischen
Unity (Georgia)
2003 U.S. Open - Men's Singles
Clarence Hammar
Erythnul
Grouse Mountain
You Must Believe in Spring
Tube Mice
Designline
GabaldÃşn
Stanley K Hornbeck
Miguel Maria N'Zau Puna
Vauxhall and I
The Young Master
Broadholme
Terragnolo
John Newman (Australian politician)
Johnny Douglas (conductor)
Berry Oakley
Francis Pemberton
3rd Shanghai International Film Festival
Harold Acton
Voices from the Sky
Pictures of Home
Dirk van Hogendorp (1761-1822)
Gus and Jaq
Christopher John Farley
Dave Hudson
2008 NCAA Men's Division I Basketball Tournament
Carol Giambalvo
Lodi High School (California)
National Technical University of Athens
Randy Bowen
Gamera 2: Attack of Legion
Ecuador
Michael Jeffery (manager)
Workers Party of the Netherlands (build-up organisation)
Michael Gallagher (translator)
UEFA Champions League 2005-06
Colleen Farrington
Robert Neal Adams
FranÃğois Jacques Boeri
Colonel By Secondary School

Marmaris
Trianon (Frankfurt am Main)
Dale Atkeson
Festive Overture (Shostakovich)
Milicent Shinn
Indira Jaising

A.2 List of Non-Entities

Acuticostites
Mitochondrial trifunctional protein
Commemorative coins of Denmark
NH RSA Title LXIII
Chamanto
OX postcode area
Security Force Auxiliaries
Formula Renault
Parliamentary representation from Buckinghamshire
Streptococcus mitis
Afflicted (band)
Armenians in Kuwait
Mepolizumab
Sports Illustrated Cover Jinx
Earl of Moray
Alanine
European Ratsnake
1006 in poetry
List of Canadian airports by location indicator: CT
Moonlander
Proper name
Neohouzeaua
Schimmel
Embryonic disk
List of host cities of the Eurovision Song Contest
Anthropoides
Nesquik
Sword-leaved Helleborine
Conformal field theory
Anta
Cocek
Administrative divisions of Chukotka Autonomous Okrug
Hedeoma pulegioides
W Ursae Majoris variable
Undulator
Cat thyme
Jenmi
Panicfire
Railroad nicknames
In My Own Time
Sensu
Fire control
Chindro
Parting tradition
Artistic License
Shadow knitting
Dirichlet algebra
Snap (dance move)
Shoshannim
County cricket
NASA Exceptional Service Medal
St. Johnstone F.C. seasons
System image
Hummock
Niederwil
List of high schools in Massachusetts
Botaniska Notiser
Reading copy
Sarcosinemia
Tramontana (sports car)
Editing Agency of Korean History
Musa (name)
Independence class aircraft carrier
Biological membrane
California Manroot
Olive (color)
300 m Standard Rifle
Restricted product

B Entities Recognized based on Categories

Following is a sample of the named entities found, grouped in their categories.

B.1 Companies

84 Lumber
AC Moore
ARAMARK
Aberdeen and Asheboro Railroad
Acme Whistles
Advanced Cell Technology
After Dark Films
Aji Ichiban
Alcon
Allens Boots
Altera
American Christian Press
American Zoetrope
Anderson Valley Brewing Company
ApS
ArcheDream
Arno Political Consultants
Ashanti Goldfields Corporation
AstraZeneca
Au
Auto AG Rothenberg
Axcom Trading Advisors
BEAM.TV
Bandwidth.com
Barclays Global Investors
Bay Networks
Belcan
Berkeley Systems
Bif Bang Pow!
Bird & Bird
Blausen Medical Communications
Bluescope Lysaght
BookFinder.com
Bowater Forest Products
BridgePort Brewing Company
British Touring Shakespeare Company
Brush Turbogenerators
Bunnpris
C venues
CJM Racing
Cabot Corporation
Calyon
Canadian Pacific hotels
Cardkey
Cary Safe Company
Celestial Digital Entertainment
Century 21
Charles Schwab Corp.
Cheshire Bus and Coach
China-based financial stocks in Hong Kong
Cincinnati Opera
Civic Hall Performing Arts Center
Cluj-Napoca Companies
Cole Haan
Commercial Aircraft Sales and Leasing
Computas AS
Container Corporation of America
Crain Communications Inc.
Cromwell Radio Group
Curves International
DC10
Dai Pai Dong
Dari Mart
De Brauw Blackstone Westbroek N.V.
Delta Faucet Company
Deutsche Bank
Digital Entertainment Network
Divine Chocolate
Dorado Wings
Drum Workshop
Dynamic packaging
EG Wrigley and Company
EarthLink
Eden Studios, Inc.
Eko guitars
Elizabeth Hurley Beach
Encore Computer
Entra Eiendom
Estar
EverBank
EyeCatcher Entertainment
Fairchild Group

Farrel Corporation
Ferrocarril General Roca
Fineos
First Second Books
Florida East Coast Railway
Forex AB
Fram
FremantleMedia
FujiGen
GAINSCO
Galaxy Communications
Gate Gourmet
Genesco Inc.
Ghana Airways
Glenmorangie
Golden Lamb Inn
Graham, Anderson, Probst & White
Great Western and Great Central Joint Railway
Group Sense PDA
Gupta Technologies
HT Motorsports
Hampshire Mall
Harman Kardon
Hay Group
Helsinki City Transport
Hideous
Hits & Favorites
Honda Atlas Cars Pakistan
Hovertravel
Hussain Industries
IBP, Inc.
ITV Digital Channels Ltd
ImageMovers Digital
Indian Railways
Inmarsat
Interceptor Micros
Interval International
Ironclad Games
J-Air
JW Marriott Hotels
Jaycar
Joffrey Ballet
Journal of Irreproducible Results
KLM Telephone
Kansas City, Pittsburgh & Gulf Railroad
Kemira
Kim Son
Klei Entertainment Inc.
Korea General Magnesia Clinker Industry Group
Kuwait Petroleum International
LXD Incorporated
Land Systems OMC
Le Coq Sportif
LendingTree
Life is Good
Lionhead Studios
List of assets owned by Time Warner
List of national and international moving associations
Lledo
Lonely Planet
Lowrance Electronics
MAN Roland
MTVX
Magna International
Manchester, South Junction and Altrincham Railway
Marcus Clark & Co.
Martin Band Instrument Company
Maurice Girodias
McKinsey & Company
Meier & Frank
Merix Corporation
Midnight Insanity
Mingxing Film Company
Mitsuwa Marketplace
Monkeystone Games
Morris & Company
Moxi
Mutual insurance
NI 43-101
Nanosight
National Orchestra Service
Nekeme Prod
Nevada Power Company
Newcastle Publishing Company
Nigerian National Petroleum Corporation
Nokia
North Eastern Railway
Norwegian types of company
Nuyorican Productions
Ocean Software

Old America Stores
Ontario Knife Company
Optus Television
Orpak
Overseas Shipholding Group
PHONE+ magazine
Pacific Publishing Company
Panic
Parry Sound Colonization Railway
Pechiney
Perceptis
Petrol Ofisi
Pic 'N' Save
Pizza Haven
Point of View, Inc.
Ports of Auckland
Presbyterian Publishing Corporation
Pro Arts Inc.
Provincial Airlines
Q-Telecom
Quicksilver Software
RPath
Raisio Group
ReactiveMicro.com
Redmonol Chemical Products Company
Renaissance Books
RheoTec Messtechnik GmbH
Riverside Methodist Hospital
Rogers Telecom
Rover
Ruskin Pottery
SBM Offshore
SNET America
SafeTV
Samsung Techwin
SaskEnergy
Schoolhouse Press
Seagull Camera
SemGroup
Seven Stories Press
Shemaroo Entertainment
Sick Room Records, LTD
Simmons Bedding Company
Skelly Oil
Smith International
Softdisk
Sonokong
Southeastern Power Administration
Spark Unlimited
SportsBooks Limited
Standard Electric Time Company
Statprobe
Stolt-Nielsen
Studio Fantasia
SunTrust Banks
Surrey Iron Railway
Symyx Technologies
TARTA
TUI Travel PLC
Tallinna Autobussikoondis
Taxijet
TeleComputing
Tembec
Texize
The Customart Press
The MathWorks
The Tabletop Group
Thomson Holidays
Time Warner
Tomioka silk mill
Towle Silversmiths
Transnational Corporation of Nigeria
Triple Canopy, Inc.
Tundra Publishing
U.S. Robotics
Ultra Electronics
United Development Company
Unsanity
Vajra Enterprises
Venray sheep companies
Victoria Express
ViroPharma
Volatile Games
WSP Group
Warner Aircraft Corporation
Weather Underground
West Coast Railway
Westnet
Wild Whirled Music
WingTips Airport Services
Woolworths
Worshipful Company of Glovers
XITEX Software
Yardbirds Home Center

Yves Saint-Laurent
Ziv Television Programs

B.2 Organizations

ANZUS
Action Palestine
Aid to Artisans
AllBusiness.com
American Association of Orthodontists
American Friends Service Committee
American Social Science Association
Animal Defenders International
Armenian Revolutionary Army
Association of Business Executives
Astrophysical Institute Potsdam
Automobile Journalists Association of Canada
Baptist Student Union
Bhaktivedanta Manor
Blue Cross and Blue Shield Association
British Association for Cemeteries in South Asia
Building society
CSTC Trenton
Canadian Association of Promotional Marketing Agencies
Canine Companions for Independence
Center for Media and Public Affairs
Charity badge
Children's Film Foundation
Churches of God General Conference
Coalition for the Good of All
Committee on Institutional Cooperation
Competitiveness Policy Council
Constantian Society
Council of Major Superiors of Women Religious
DAIA
Death squad
Diamond Sangha
EC-SAR
Education Conservancy
Engineers for a Sustainable World
European Association of Conservatoires
European and Mediterranean Plant Protection Organization
Famine Early Warning Systems Network
Film Unit
Forsaken
French Defence Health service
Gawad Kalinga
Girl Guides Association of Papua New Guinea
Got Questions
Guidelines International Network
Harvard-Radcliffe Science Fiction Association
Hindu Makkal Katchi
Howard Brown Health Center
IPIC
Independent Task Force on North America
Institute in Basic Life Principles
International Accounting Standards Committee
International Colour Authority
International Football Association Board
International Progress Organization
International Yoga Federation
Islamic Mission of Belize
Japan Baptist Association
John Aspinall Foundation
Kashi Mutt
Kobayashi aikido
Lake View Citizens' Council
Legion of Doom
List of Aikido organizations
List of fictional companies
London Club
Magician Alliance of Eastern States
MassEquality
Merit School of Music
Minnesota Zen Center
Muddy York Rugby Football Club
NCPAD
National Association of Military Marching Bands
National Council of Resistance of Iran
National Lesbian and Gay Journalists Asso-

ciation
National Union of South African Students
New England Research Institutes
Nippon Foundation
Norwegian Maritime Directorate
Odinic Rite
OpenTravel Alliance
Orpheum Foundation for the Advancement of Young Soloists
Pakistan Boy Scouts Association
Peace Society
Philalethes Society
Political Research Associates
Program for Appropriate Technology in Health
Quackwatch
Republican Conference Chairman of the United States Senate
Rodobrana
Royal Order of Scotland
SPQ Libre
Self-Realization Fellowship
Sigma Theta Epsilon
Society for Electro-Acoustic Music in the United States
Soroptimist
Sporting Arms and Ammunition Manufacturers' Institute
Student Environmental Action Coalition
Swedish Film Institute
Taxpayer groups
The Banyan
The Girl Guides Association of Antigua and Barbuda
The Order
The Waffle
Transportation Alternatives
UFORM
Union of International Associations
United States National Karate Association
Vaccine and Infectious Disease Organization
Vision America
Wayne RESA
Wireless Toronto
World Buddhist Forum
World Taiwanese Congress
Young Men's Institute

B.3 People

"Hungry" Charles Hardy
Abdur Razzak
Ahmet Zappa
Alan Brinkley
Albert and David Maysles
Alex Grammas
Alexi Giannoulias
Aliza Olmert
Amber MacArthur
Andrew Howe
Anton Villatoro
Arild Andersen
Arturo Torres
Avery Cardoza
Barbara Mertz
Becky Morgan
Beverlei Brown
Bill Schwab
Blu Greenberg
Bob Wolff
Brad Childress
Brent Patterson
Brian Price
Bruce Reid
Carl Hewitt
Carmine Boal
Cathy Hughes
Charles E. Barkley
Chase Daniels
Chris Burke
Chris Smith
Christophe Bordeau
Cindy O'Callaghan
Clifford Ray
Conrad Brooks
Craig Sager
D. Ray Perdue Jr.
Dan Gillespie Sells
Daniel Kaluuya
Danny Strong
Dashon Goldson
David Atherton
David Giffin
David Meyer
David Ushery
Deborah Gordon
Dennis K. Villa
Dimitar Stilianov
Don Carter
Donovan Patton
Drew Coleman
Eberhard Weise
Edmund Purdom
Eitan Cabel
Ella Tripp
Emmanuel Lubezki
Eric Rupe

Erwin Schild
Ewan McCray
Felipe Baloy
Floris Jansen
Frank Broome
Freaky Flow
Fuzzy Zoeller
Gary Anderson
Geert Versnick
George Gao
Gerald Sibon
Gil da Cruz Trindade
Glenn Kaiser
Graham Day
Gregory C. Farrington
Guy Whittall
Hank Aaron
Harry Fowler
Heinrich Mussinghoff
Herb Grubel
Holly Davidson
Hugues Claude Pissarro
Ian Sample
Isolde Kostner
J. Stuart Perkins
Jacob Smith
James Blaylock
James O'Connor
Jared Boice
Javid Hussain
Jean-Jacques Burnel
Jeff Sagarin
Jeon Kwang-cheol
Jim Doyle
Jimmy Dixon
Jodi Santamaria
Joel Dreessen
John Branney
John Gardiner
John Sabini
Johnny Kerr
Jonathan Kerrigan
Julianne Baird
Justin Wheatley
Kang Soo Jin
Katalin Szili
Kaylynn
Kelly Overton
Kenneth Schellenberger
Kevin L. Bryant
Kim Jagtiani
Ko Jong-Soo
Kunio Kitamura
Lance Davids
Laura Freixas
Lee Blackburn
Leo Hayden
Lew Krausse Jr.
Lindsay Frost
Logan Vander Velden
Lowitja O'Donoghue
MC Romeo
Malcolm Boyden
Marc Gicquel
Marcus Stephen
Marie Plourde
Mark Blundell
Mark Ormrod
Marshall Faulk
Marty Feldman
Masashi Nakayama
Matt Stewart
Mauricio de Sousa
Mel Machin
Michael Blaudzun
Michael Johnson
Michael Stegmayer
MichÃČÂĺle Jacot
Mike Deodato
Mike Stahr
Mohammad Reza Mamani
Moshe Ohayon
Nacanieli Seru
Natalio Lorenzo Poquet
Neil Nunes
Nick Johnson
Niilo Halonen
Padraig Parkinson
Pat Sobeski
Pattie Boyd
Paul Kehoe
Paul de Casteljau
Per WikstrÃŕÂ£Â¡m
Peter G. Tsouras
Peter Staples
Philip Carlo
Piet Keizer
Prosper Avril
Rafael Palmeiro
Randall Godfrey
Ray Williams
Renaldas Seibutis
Richard A. Pittman
Richard O. Spertzel
Ricky Steamboat
Robert AhMat
Roel Luynenburg
Ron Allen
Rory McCarthy
Ryan Gosling
Sajib Miah
Sammy Lee
Sarah Huck
Scott Maslen
Seiji Osaka
Shahid Israr
Shawn Stasiak
Shona Moller
Simon Mrashani
Sonja Bennett
Stephen Lodge
Steve Kariya
Steven Rathman
Sulley Muntari
Takako Katou
Tatiana Poutchek
Terry Bickers
Thom Fitzgerald
Tiffany Brissette
Timothy R. Ferguson
Tom Dine
Tony Kendall
Travis Diener
Ty Esler
Valentin Simion
Vic Bubas
Vincent Ribeton
Warren Munson
Wilfried Nelissen
William Prochnau
Wu Shih-Hsih

Yoann Lachor Zintis Ekmanis
Yuval Yairi