Mining Wiki Resources for Multilingual Named Entity Recognition

Alexander E. Richman
Department of Defense
Washington, DC 20310
arichman@psualum.com

Patrick Schone
Department of Defense
Fort George G. Meade, MD 20755
pjschon@tycho.ncsc.mil

Abstract

In this paper, we describe a system by which the multilingual characteristics of Wikipedia can be utilized to annotate a large corpus of text with Named Entity Recognition (NER) tags requiring minimal human intervention and no linguistic expertise. This process, though of value in languages for which resources exist, is particularly useful for less commonly taught languages. We show how the Wikipedia format can be used to identify possible named entities and discuss in detail the process by which we use the Category structure inherent to Wikipedia to determine the named entity type of a proposed entity. We further describe the methods by which English language data can be used to bootstrap the NER process in other languages. We demonstrate the system by using the generated corpus as training sets for a variant of BBN's Identifinder in French, Ukrainian, Spanish, Polish, Russian, and Portuguese, achieving overall F-scores as high as 84.7% on independent, human-annotated corpora, comparable to a system trained on up to 40,000 words of human-annotated newswire.

1 Introduction

Named Entity Recognition (NER) has long been a major task of natural language processing. Most of the research in the field has been restricted to a few languages and almost all methods require substantial linguistic expertise, whether creating a rule-based technique specific to a language or manually annotating a body of text to be used as a training set for a statistical engine or machine learning.

In this paper, we focus on using the multilingual Wikipedia (wikipedia.org) to automatically create an annotated corpus of text in any given language, with no linguistic expertise required on the part of the user at run-time (and only English knowledge required during development). The expectation is that for any language in which Wikipedia is sufficiently well-developed, a usable set of training data can be obtained with minimal human intervention. As Wikipedia is constantly expanding, it follows that the derived models are continually improved and that increasingly many languages can be usefully modeled by this method.

In order to make sure that the process is as language-independent as possible, we declined to make use of any non-English linguistic resources outside of the Wikimedia domain (specifically, Wikipedia and the English language Wiktionary (en.wiktionary.org)). In particular, we did not use any semantic resources such as WordNet or part of speech taggers. We used our automatically annotated corpus along with an internally modified variant of BBN's IdentiFinder (Bikel et al., 1999), specifically modified to emphasize fast text processing, called "PhoenixIDF," to create several language models that could be tested outside of the Wikipedia framework. We built on top of an existing system, and left existing lists and tables intact. Depending on language, we evaluated our derived models against human or machine annotated data sets to test the system.

2 Wikipedia

2.1 Structure

Wikipedia is a multilingual, collaborative encyclopedia on the Web which is freely available for research purposes. As of October 2007, there were over 2 million articles in English, with versions available in 250 languages. This includes 30 languages with at least 50,000 articles and another 40 with at least 10,000 articles. Each language is available for download (download.wikimedia.org) in a text format suitable for inclusion in a database. For the remainder of this paper, we refer to this format.

Proceedings of ACL-08: HLT, pages 1–9,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

Within Wikipedia, we take advantage of five major features:

- *Article links,* links from one article to another of the same language;
- *Category links*, links from an article to special "Category" pages;
- *Interwiki links*, links from an article to a presumably equivalent, article in another language;
- *Redirect pages*, short pages which often provide equivalent names for an entity; and
- *Disambiguation pages*, a page with little content that links to multiple similarly named articles.

The first three types are collectively referred to as *wikilinks*.

A typical sentence in the database format looks like the following:

"Nescopeck Creek is a [[tributary]] of the [[North Branch Susquehanna River]] in [[Luzerne County, Pennsylvania|Luzerne County]]."

The double bracket is used to signify wikilinks. In this snippet, there are three articles links to English language Wikipedia pages, titled "Tributary," "North Branch Susquehanna River," and "Luzerne County, Pennsylvania." Notice that in the last link, the phrase preceding the vertical bar is the name of the article, while the following phrase is what is actually displayed to a visitor of the webpage.

Near the end of the same article, we find the following representations of Category links: [[Category:Luzerne County, Pennsylvania]], [[Category:Rivers of Pennsylvania]], {{Pennsylvania-geo-stub}}. The first two are direct links to Category pages. The third is a link to a Template, which (among other things) links the article to "Category:Pennsylvania geography stubs". We will typically say that a given entity *belongs to* those categories to which it is linked in these ways.

The last major type of wikilink is the link between different languages. For example, in the Turkish language article "Kanuni Sultan Süleyman" one finds a set of links including [[en:Suleiman the Magnificent]] and [[ru:Сулейман I]]. These represent links to the English language article "Suleiman the Magnificent" and the Russian language article "Сулейман I." In almost all cases, the articles linked in this manner represent articles on the same subject.

A redirect page is a short entry whose sole purpose is to direct a query to the proper page. There are a few reasons that redirect pages exist, but the primary purpose is exemplified by the fact that "USA" is an entry which redirects the user to the page entitled "United States." That is, in the vast majority of cases, redirect pages provide another name for an entity.

A disambiguation page is a special article which contains little content but typically lists a number of entries which might be what the user was seeking. For instance, the page "Franklin" contains 70 links, including the singer "Aretha Franklin," the town "Franklin, Virginia," the "Franklin River" in Tasmania, and the cartoon character "Franklin (Peanuts)." Most disambiguation pages are in Category:Disambiguation or one of its subcategories.

2.2 Related Studies

Wikipedia has been the subject of a considerable amount of research in recent years including Gabrilovich and Markovitch (2007), Strube and Ponzetto (2006), Milne et al. (2006), Zesch et al. (2007), and Weale (2007). The most relevant to our work are Kazama and Torisawa (2007), Toral and Muñoz (2006), and Cucerzan (2007). More details follow, but it is worth noting that all known prior results are fundamentally monolingual, often developing algorithms that can be adapted to other languages pending availability of the appropriate semantic resource. In this paper, we emphasize the use of links between articles of different languages, specifically between English (the largest and best linked Wikipedia) and other languages.

Toral and Muñoz (2006) used Wikipedia to create lists of named entities. They used the first sentence of Wikipedia articles as likely definitions of the article titles, and used them to attempt to classify the titles as people, locations, organizations, or none. Unlike the method presented in this paper, their algorithm relied on WordNet (or an equivalent resource in another language). The authors noted that their results would need to pass a manual supervision step before being useful for the NER task, and thus did not evaluate their results in the context of a full NER system.

Similarly, Kazama and Torisawa (2007) used Wikipedia, particularly the first sentence of each article, to create lists of entities. Rather than building entity dictionaries associating words and phrases to the classical NER tags (PERSON, LOCATION, etc.) they used a noun phrase following forms of the verb “to be” to derive a label. For example, they used the sentence “Franz Fischler ... is an Austrian politician” to associate the label “politician” to the surface form “Franz Fischler.” They proceeded to show that the dictionaries generated by their method are useful when integrated into an NER system. We note that their technique relies upon a part of speech tagger, and thus was not appropriate for inclusion as part of our non-English system.

Cucerzan (2007), by contrast to the above, used Wikipedia primarily for Named Entity Disambiguation, following the path of Bunescu and Paşca (2006). As in this paper, and unlike the above mentioned works, Cucerzan made use of the explicit Category information found within Wikipedia. In particular, Category and related list-derived data were key pieces of information used to differentiate between various meanings of an ambiguous surface form. Unlike in this paper, Cucerzan did not make use of the Category information to identify a given entity as a member of any particular class. We also note that the NER component was not the focus of the research, and was specific to the English language.

3 Training Data Generation

3.1 Initial Set-up and Overview

Our approach to multilingual NER is to pull back the decision-making process to English whenever possible, so that we could apply some level of linguistic expertise. In particular, by focusing on only one language, we could take maximum advantage of the Category structure, something very difficult to do in the general multilingual case.

For computational feasibility, we downloaded various language Wikipedias and the English language Wiktionary in their text (.xml) format and stored each language as a table within a single MySQL database. We only stored the title, id number, and body (the portion between the <TEXT> and </TEXT> tags) of each article.

We elected to use the ACE Named Entity types PERSON, GPE (Geo-Political Entities), ORGANIZATION, VEHICLE, WEAPON, LOCATION, FACILITY, DATE, TIME, MONEY, and PERCENT. Of course, if some of these types were not marked in an existing corpus or not needed for a given purpose, the system can easily be adapted.

Our goal was to automatically annotate the text portion of a large number of non-English articles with tags like <ENAMEX TYPE=“GPE”>Place Name</ENAMEX> as used in MUC (Message Understanding Conference). In order to do so, our system first identifies words and phrases within the text that might represent entities, primarily through the use of wikilinks. The system then uses category links and/or interwiki links to associate that phrase with an English language phrase or set of Categories. Finally, it determines the appropriate type of the English language data and assumes that the original phrase is of the same type.

In practice, the English language categorization should be treated as one-time work, since it is identical regardless of the language model being built. It is also the only stage of development at which we apply substantial linguistic knowledge, even of English.

In the sections that follow, we begin by showing how the English language categorization is done. We go on to describe how individual non-English phrases are associated with English language information. Next, we explain how possible entities are initially selected. Finally, we discuss some optional steps as well as how and why they could be used.

3.2 English Language Categorization

For each article title of interest (specifically excluding Template pages, Wikipedia admistrative pages, and articles whose title begins with “List of”), we extracted the categories to which that entry was assigned. Certainly, some of these category assignments are much more useful than others

For instance, we would expect that any entry in “Category:Living People” or “Category:British Lawyers” will refer to a person while any entry in “Category:Cities in Norway” will refer to a GPE. On the other hand, some are entirely unhelpful, such as “Category:1912 Establishments” which includes articles on Fenway Park (a facility), the Republic of China (a GPE), and the Better Business Bureau (an organization). Other categories can reliably be used to determine that the article does not refer to a named entity, such as “Category:Endangered species.” We manually derived a relatively small set of key phrases, the most important of which are shown in Table 1.

Table 1: Some Useful Key Category Phrases

PERSON	"People by", "People in", "People from", "Living people", "births", "deaths", "by occupation", "Surname", "Given names", "Biography stub", "human names"
ORG	"Companies", "Teams", "Organizations", "Businesses", "Media by", "Political parties", "Clubs", "Advocacy groups", "Unions", "Corporations", "Newspapers", "Agencies", "Colleges", "Universities" , "Legislatures", "Company stub", "Team stub", "University stub", "Club stub"
GPE	"Cities", "Countries", "Territories", "Counties", "Villages", "Municipalities", "States" (not part of "United States"), "Republics", "Regions", "Settlements"
DATE	"Days", "Months", "Years", "Centuries"
NONE	"Lists", "List of", "Wars", "Incidents"

For each article, we searched the category hierarchy until a threshold of reliability was passed or we had reached a preset limit on how far we would search.

For example, when the system tries to classify "Jacqueline Bhabha," it extracts the categories "British Lawyers," "Jewish American Writers," and "Indian Jews." Though easily identifiable to a human, none of these matched any of our key phrases, so the system proceeded to extract the second order categories "Lawyers by nationality," "British legal professionals," "American writers by ethnicity," "Jewish writers," "Indian people by religion," and "Indian people by ethnic or national origin" among others. "People by" is on our key phrase list, and the two occurrences passed our threshold, and she was then correctly identified.

If an article is not classified by this method, we check whether it is a disambiguation page (which often are members solely of "Category:Disambiguation"). If it is, the links within are checked to see whether there is a dominant type. For instance, the page "Amanda Foreman" is a disambiguation page, with each link on the page leading to an easily classifiable article.

Finally, we use Wiktionary, an online collaborative dictionary, to eliminate some common nouns. For example, "Tributary" is an entry in Wikipedia which would be classified as a Location if viewed solely by Category structure. However, it is found as a common noun in Wiktionary, overruling the category based result.

3.3 Multilingual Categorization

When attempting to categorize a non-English term that has an entry in its language's Wikipedia, we use two techniques to make a decision based on English language information. First, whenever possible, we find the title of an associated English language article by searching for a wikilink beginning with "en:". If such a title is found, then we categorize the English article as shown in Section 3.2, and decide that the non-English title is of the same type as its English counterpart. We note that links to/from English are the most common interlingual wikilinks.

Of course, not all articles worldwide have English equivalents (or are linked to such even if they do exist). In this case, we attempt to make a decision based on Category information, associating the categories with their English equivalents, when possible. Fortunately, many of the most useful categories have equivalents in many languages.

For example, the Breton town of Erquy has a substantial article in the French language Wikipedia, but no article in English. The system proceeds by determining that Erquy belongs to four French language categories: "Catégorie:Commune des Côtes-d'Armor," "Catégorie:Ville portuaire de France," "Catégorie:Port de plaisance," and "Catégorie:Station balnéaire." The system proceeds to associate these, respectively, with "Category:Communes of Côtes-d'Armor," UNKNOWN, "Category:Marinas," and "Category:Seaside resorts" by looking in the French language pages of each for wikilinks of the form [[en:...]].

The first is a subcategory of "Category:Cities, towns and villages in France" and is thus easily identified by the system as a category consisting of entities of type GPE. The other two are ambiguous categories (facility and organization elements in addition to GPE). Erquy is then determined to be a GPE by majority vote of useful categories.

We note that the second French category actually has a perfectly good English equivalent (Category:Port cities and towns in France), but no one has linked them as of this writing. We also note that the ambiguous categories are much more GPE-oriented in French. The system still makes the correct decision despite these factors.

We do not go beyond the first level categories or do any disambiguation in the non-English case. Both are avenues for future improvement.

3.4 The Full System

To generate a set of training data in a given language, we select a large number of articles from its Wikipedia (50,000 or more is recommended, when possible). We prepare the text by removing external links, links to images, category and interlingual links, as well as some formatting. The main processing of each article takes place in several stages, whose primary purposes are as follows:

- The first pass uses the explicit article links within the text.
- We then search an associated English language article, if available, for additional information.
- A second pass checks for multi-word phrases that exist as titles of Wikipedia articles.
- We look for certain types of person and organization instances.
- We perform additional processing for alphabetic or space-separated languages, including a third pass looking for single word Wikipedia titles.
- We use regular expressions to locate additional entities such as numeric dates.

In the first pass, we attempt to replace all wikilinks with appropriate entity tags. We assume at this stage that any phrase identified as an entity at some point in the article will be an entity of the same type throughout the article, since it is common for contributors to make the explicit link only on the first occasion that it occurs. We also assume that a phrase in a bold font within the first 100 characters is an equivalent form of the title of the article as in this start of the article on Erquy: "**Erquy** (**Erge-ar-Mor** en breton, **Erqi** en gallo)". The parenthetical notation gives alternate names in the Breton and Gallo languages. (In Wiki database format, bold font is indicated by three apostrophes in succession.)

If the article has an English equivalent, we search that article for wikilinked phrases as well, on the assumption that both articles will refer to many of the same entities. As the English language Wikipedia is the largest, it frequently contains explicit references to and articles on secondary people and places mentioned, but not linked, within a given non-English article. After this point, the text to be annotated contains no Wikipedia specific information or formatting.

In the second pass, we look for strings of 2 to 4 words which were not wikilinked but which have Wikipedia entries of their own or are partial matches to known people and organizations (i.e. "Mary Washington" in an article that contains "University of Mary Washington"). We require that each such string contains something other than a lower case letter (when a language does not use capitalization, nothing in that writing system is considered to be lower case for this purpose). When a word is in more than one such phrase, the longest match is used.

We then do some special case processing. When an organization is followed by something in parentheses such as <ENAMEX TYPE="ORGANIZATION">Maktab al-Khadamāt</ENAMEX> (MAK), we hypothesize that the text in the parentheses is an alternate name of the organization. We also looked for unmarked strings of the form X.X. followed by a capitalized word, where X represents any capital letter, and marked each occurrence as a PERSON.

For space-separated or alphabetic languages, we did some additional processing at this stage to attempt to identify more names of people. Using a list of names derived from Wiktionary (Appendix:Names) and optionally a list derived from Wikipedia (see Section 3.5.1), we mark possible parts of names. When two or more are adjacent, we mark the sequence as a PERSON. Also, we fill in partial lists of names by assuming single non-lower case words between marked names are actually parts of names themselves. That is, we would replace <ENAMEX TYPE="PERSON">Fred Smith</ENAMEX>, Somename <ENAMEX TYPE="PERSON">Jones </ENAMEX> with <ENAMEX TYPE="PERSON"> Fred Smith</ENAMEX>, <ENAMEX TYPE= "PERSON"> Somename Jones</ENAMEX>. At this point, we performed a third pass through the article. We marked all non-lower case single words which had their own Wikipedia entry, were part of a known person's name, or were part of a known organization's name.

Afterwards, we used a series of simple, language-neutral regular expressions to find additional TIME, PERCENT, and DATE entities such as "05:30" and "12-07-05". We also executed code that included quantities of money within a NUMEX tag, as in converting 500 <NUMEX TYPE="MONEY">USD</NUMEX> into <NUMEX TYPE="MONEY">500 USD</NUMEX>.

3.5 Optional Processing

3.5.1 Recommended Additions

All of the above could be run with almost no understanding of the language being modeled (knowing whether the language was space-separated and whether it was alphabetic or character-based were the only things used). However, for most languages, we spent a small amount of time (less than one hour) browsing Wikipedia pages to improve performance in some areas.

We suggest compiling a small list of stop words. For our purposes, the determiners and the most common prepositions are sufficient, though a longer list could be used for the purpose of computational efficiency.

We also recommend compiling a list of number words as well as compiling a list of currencies, since they are not capitalized in many languages, and may not be explicitly linked either. Many languages have a page on ISO 4217 which contains all of the currency information, but the format varies sufficiently from language to language to make automatic extraction difficult. Together, these allow phrases like this (taken from the French Wikipedia) to be correctly marked in its entirety as an entity of type MONEY: "25 millions de dollars."

If a language routinely uses honorifics such as Mr. and Mrs., that information can also be found quickly. Their use can lead to significant improvements in PERSON recognition.

During preprocessing, we typically collected a list of people names automatically, using the entity identification methods appropriate to titles of Wikipedia articles. We then used these names along with the Wiktionary derived list of names during the main processing. This does introduce some noise as the person identification is not perfect, but it ordinarily increases recall by more than it reduces precision.

3.5.2 Language Dependent Additions

Our usual, language-neutral processing only considers wikilinks within a single article when determining the type of unlinked words and phrases. For example, if an article included the sentence "The [[Delaware River|Delaware]] forms the boundary between [[Pennsylvania]] and [[New Jersey]]", our system makes the assumption that every occurrence of the unlinked word "Delaware" appearing in the same article is also referring to the river and thus mark it as a LOCATION.

For some languages, we preferred an alternate approach, best illustrated by an example: The word "Washington" without context could refer to (among others) a person, a GPE, or an organization. We could work through all of the explicit wikilinks in all articles (as a preprocessing step) whose surface form is Washington and count the number pointing to each. We could then decide that every time the word Washington appears without an explicit link, it should be marked as its most common type. This is useful for the Slavic languages, where the nominative form is typically used as the title of Wikipedia articles, while other cases appear frequently (and are rarely wikilinked).

At the same time, we can do a second type of preprocessing which allows more surface forms to be categorized. For instance, imagine that we were in a Wikipedia with no article or redirect associated to "District of Columbia" but that someone had made a wikilink of the form [[Washington|District of Columbia]]. We would then make the assumption that for all articles, District of Columbia is of the same type as Washington.

For less developed wikipedias, this can be helpful. For languages that have reasonably well developed Wikipedias and where entities rarely, if ever, change form for grammatical reasons (such as French), this type of preprocessing is virtually irrelevant. Worse, this processing is definitely not recommended for languages that do not use capitalization because it is not unheard of for people to include sections like: "The [[Union Station|train station]] is located at ..." which would cause the phrase "train station" to be marked as a FACILITY each time it occurred. Of course, even in languages with capitalization, "train station" would be marked incorrectly in the article in which the above was located, but the mistake would be isolated, and should have minimal impact overall.

4 Evaluation and Results

After each data set was generated, we used the text as a training set for input to PhoenixIDF. We had three human annotated test sets, Spanish, French and Ukrainian, consisting of newswire. When human annotated sets were not available, we held out more than 100,000 words of text generated by our wiki-mining process to use as a test set. For the above languages, we included wiki test sets for comparison purposes. We will give our results as F-scores in the Overall, DATE, GPE, ORGANIZATION, and PERSON categories using the scoring metric in (Bikel et. al, 1999). The other ACE categories are much less common, and contribute little to the overall score.

4.1 Spanish Language Evaluation

The Spanish Wikipedia is a substantial, well-developed Wikipedia, consisting of more than 290,000 articles as of October 2007. We used two test sets for comparison purposes. The first consists of 25,000 words of human annotated newswire derived from the ACE 2007 test set, manually modified to conform to our extended MUC-style standards. The second consists of 335,000 words of data generated by the Wiki process held-out during training.

Table 2: Spanish Results

F (prec. / recall)	**Newswire**	**Wiki test set**
ALL	**.827** (.851 / .805)	.846 (.843 / .848)
DATE	.912 (.861 / .970)	.925 (.918 / .932)
GPE	.877 (.914 / .843)	.877 (.886 / .868)
ORG	.629 (.681 / .585)	.701 (.703 / .698)
PERSON	.906 (.921 / .892)	.821 (.810 / .833)

There are a few particularly interesting results to note. First, because of the optional processing, recall was boosted in the PERSON category at the expense of precision. The fact that this category scores higher against newswire than against the wiki data suggests that the not-uncommon, but isolated, occurrences of non-entities being marked as PERSONs in training have little effect on the overall system. Contrarily, we note that deletions are the dominant source of error in the ORGANIZATION category, as seen by the lower recall. The better performance on the wiki set seems to suggest that either Wikipedia is relatively poor in Organizations or that PhoenixIDF underperforms when identifying Organizations relative to other categories or a combination.

An important question remains: "How do these results compare to other methodologies?" In particular, while we can get these results for free, how much work would traditional methods require to achieve comparable results?

To attempt to answer this question, we trained PhoenixIDF on additional ACE 2007 Spanish language data converted to MUC-style tags, and scored its performance using the same set of newswire. Evidently, comparable performance to our Wikipedia derived system requires between 20,000 and 40,000 words of human-annotated newswire. It is worth noting that Wikipedia itself is not newswire, so we do not have a perfect comparison.

Table 3: Traditional Training

~ Words of Training	Overall F-score
3500	.746
10,000	.760
20,000	**.807**
40,000	**.847**

4.2 French Language Evaluation

The French Wikipedia is one of the largest Wikipedias, containing more than 570,000 articles as of October 2007. For this evaluation, we have 25,000 words of human annotated newswire (*Agence France Presse*, 30 April and 1 May 1997) covering diverse topics. We used 920,000 words of Wiki-derived data for the second test.

Table 4: French Results

F (prec. / recall)	**Newswire**	**Wiki test set**
ALL	**.847** (.877 / .819)	.844 (.847 / .840)
DATE	.921 (.897 / .947)	.910 (.888 / .934)
GPE	.907 (.933 / .882)	.868 (.889 / .849)
ORG	.700 (.794 / .625)	.718 (.747 / .691)
PERSON	.880 (.874 / .885)	.823 (.818 / .827)

The overall results seem comparable to the Spanish, with the slightly better overall performance likely correlated to the somewhat more developed Wikipedia. We did not have sufficient quantities of annotated data to run a test of the traditional methods, but Spanish and French are sufficiently similar languages that we expect this model is comparable to one created with about 40,000 words of human-annotated data.

4.3 Ukrainian Language Evaluation

The Ukrainian Wikipedia is a medium-sized Wikipedia with 74,000 articles as of October 2007. Also, the typical article is shorter and less well-linked to other articles than in the French or Spanish versions. Moreover, entities tend to appear in many surface forms depending on case, leading us to expect somewhat worse results. In the Ukrainian case, the newswire consisted of approximately 25,000 words from various online news sites covering primarily political topics. We also held out around 395,000 words for testing. We were also able to run a comparison test as in Spanish.

Table 5: Ukrainian Results

F (prec. / recall)	**Newswire**	**Wiki test set**
ALL	**.747** (.863 / .649)	.807 (.809 / .806)
DATE	.780 (.759 / .803)	.848 (.842 / .854)
GPE	.837 (.833 / .841)	.887 (.901 / .874)
ORG	.585 (.800 / .462)	.657 (.678 / .637)
PERSON	.764 (.899 / .664)	.690 (.675 / .706)

Table 6: Traditional Training

~ Words of Training	Overall F-score
5000	.662
10,000	.692
15,000	**.740**
20,000	**.761**

The Ukrainian newswire contained a much higher proportion of organizations than the French or Spanish versions, contributing to the overall lower score. The Ukrainian language Wikipedia itself contains very few articles on organizations relative to other types, so the distribution of entities of the two test sets are quite different. We also see that the Wiki-derived model performs comparably to a model trained on 15-20,000 words of human-annotated text.

4.4 Other Languages

For Portuguese, Russian, and Polish, we did not have human annotated corpora available for testing. In each case, at least 100,000 words were held out from training to be used as a test set. It seems safe to suppose that if suitable human-annotated sets were available for testing, the PERSON score would likely be higher, and the ORGANIZATION score would likely be lower, while the DATE and GPE scores would probably be comparable.

Table 7: Other Language Results

F-score	**Polish**	**Portuguese**	**Russian**
ALL	**.859**	**.804**	**.802**
DATE	.891	.861	.822
GPE	.916	.826	.867
ORG	.785	.706	.712
PERSON	.836	.802	.751

5 Conclusions

In conclusion, we have demonstrated that Wikipedia can be used to create a Named Entity Recognition system with performance comparable to one developed from 15-40,000 words of human-annotated newswire, while not requiring any linguistic expertise on the part of the user. This level of performance, usable on its own for many purposes, can likely be obtained currently in 20-40 languages, with the expectation that more languages will become available, and that better models can be developed, as Wikipedia grows.

Moreover, it seems clear that a Wikipedia-derived system could be used as a supplement to other systems for many more languages. In particular, we have, for all practical purposes, embedded in our system an automatically generated entity dictionary.

In the future, we would like to find a way to automatically generate the list of key words and phrases for useful English language categories. This could implement the work of Kazama and Torisawa, in particular. We also believe performance could be improved by using higher order non-English categories and better disambiguation. We could also experiment with introducing automatically generated lists of entities into PhoenixIDF directly. Lists of organizations might be particularly useful, and “List of” pages are common in many languages.

References

Bikel, D., R. Schwartz, and R. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 211-31.

Bunescu, R and M. Paşca. 2006. Using Encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, 9-16.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP/CoNLL*, 708-16.

Gabrilovitch, E. and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, 1606-11.

Gabrilovitch, E. and S. Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *Proceedings of AAAI*, 1301-06.

Gabrilovitch, E. and S. Markovitch. 2005. Feature generation for text categorization using world knowledge. In *Proceedings of IJCAI*, 1048-53.

Kazama, J. and K. Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP/CoNLL*, 698-707.

Milne, D., O. Medelyan and I. Witten. 2006. Mining domain-specific thesauri from Wikipedia: a case study. *Web Intelligence 2006*, 442-48

Strube, M. and S. P. Ponzeto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of AAAI*, 1419-24.

Toral, A. and R. Muñoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of EACL*, 56-61.

Weale, T. 2006. Using Wikipedia categories for document classification. *Ohio St. University, preprint*.

Zesch, T., I. Gurevych and M. Mühlhäuser. 2007. Analyzing and accessing Wikipedia as a lexical semantic resource. In *Proceedings of GLDV*, 213-21.