Cross-Language Entity Linking

Paul McNamee and **James Mayfield**
HLTCOE & Applied Physics Laboratory
Johns Hopkins University
{paul.mcnamee,james.mayfield}@jhuapl.edu

Dawn Lawrie
Loyola University Maryland
lawrie@cs.loyola.edu

Douglas W. Oard
iSchool & UMIACS
University of Maryland, College Park
oard@umd.edu

David Doermann
UMIACS
University of Maryland, College Park
doermann@umiacs.umd.edu

Abstract

There has been substantial recent interest in aligning mentions of named entities in unstructured texts to knowledge base descriptors, a task commonly called *entity linking*. This technology is crucial for applications in knowledge discovery and text data mining. This paper presents experiments in the new problem of *cross-language entity linking*, where documents and named entities are in a different language than that used for the content of the reference knowledge base. We have created a new test collection to evaluate cross-language entity linking performance in twenty-one languages. We present experiments that examine issues such as: the importance of transliteration; the utility of cross-language information retrieval; and, the potential benefit of multilingual named entity recognition. Our best model achieves performance which is 94% of a strong monolingual baseline.

1 Introduction

Entity Linking involves aligning a textual mention of a named entity to the entry in a knowledge base (KB) that represents the mentioned entity, if it is present. The problem has two main complicating features: entities can be referred to using multiple name variants (*e.g.,* aliases or misspellings); and several entities can share the same name (*e.g.,* many people are named María Sánchez). Applications of entity linking include linking patient health records from separate hospitalizations, maintaining personal credit files, preventing identity crimes, and supporting law enforcement.

Starting in 2009 the NIST Text Analysis Conference (TAC) began conducting evaluations of technologies for knowledge base population (KBP). Systems addressing the entity linking subtask take as input a name string from a document and produce as output the knowledge base node, if any, corresponding to the mentioned entity. This capability is vital for knowledge discovery; without it, extracted information cannot be properly inserted in the correct KB node. We follow the TAC-KBP problem formulation and use its reference knowledge base, which was derived from a 2008 snapshot of English Wikipedia. In the present work our focus is on person entities. We seek to develop and evaluate technologies for matching foreign language names to the appropriate knowledge base descriptor (or *kbid*) in the English KB.

To support this research we created what we believe to be the first cross-language entity linking test collection. Our dataset includes twenty-one languages in addition to English, and covers five writing systems. Compared to the problem of monolingual (English) entity linking, a solution to the cross-language variant requires both a method to match foreign names to their English equivalents, and a way to compare contextual features from the non-English source document with contextual information about known entities stored in the KB. Figure 1 illustrates the process.

The rest of this paper is structured as follows. In Section 2 we discuss related work in entity linking and cross-language name matching. In Section 3 we present our approach to monolingual entity linking and describe the adaptations that are required to address the cross-language problem. Section 4 discusses the TAC-KBP evaluation and the construction of our test collection. Sections 5 and 6 present experiments exploring the effects of transliteration and cross-language content matching on the problem. Section 7 summarizes the main contributions of this work.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 255–263,
Chiang Mai, Thailand, November 8 – 13, 2011.

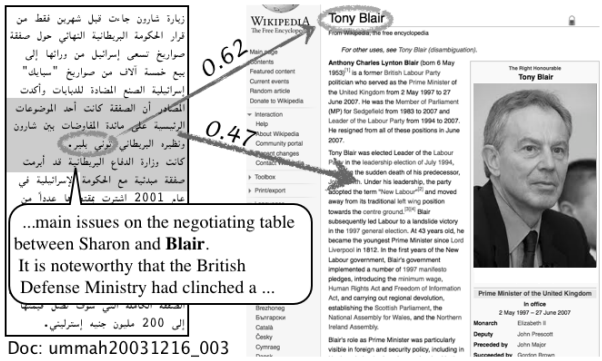

Figure 1: Linking an Arabic query referring to Tony Blair to a Wikipedia-derived KB using name matching and context matching features.

⟨entity wiki_title="Tony_Blair" type="PER" id="E0481157" name="Tony Blair"⟩
⟨facts class="Infobox Prime Minister"⟩
⟨fact name="honorific-prefix"⟩The Right Honourable⟨/fact⟩
⟨fact name="name"⟩Tony Blair⟨/fact⟩
⟨fact name="order"⟩Prime Minister of the United Kingdom⟨/fact⟩
⟨fact name="monarch"⟩⟨link entity_id="E0699345"⟩Elizabeth II⟨/link⟩⟨/fact⟩
⟨fact name="predecessor"⟩⟨link entity_id="E0614790"⟩John Major⟨/link⟩⟨/fact⟩
⟨fact name="successor"⟩⟨link entity_id="E0455080"⟩Gordon Brown⟨/link⟩⟨/fact⟩
⟨fact name="birth_date"⟩6 May 1953 (1953-05-06) (age56)⟨/fact⟩
⟨fact name="spouse"⟩⟨link entity_id="E0629105"⟩Cherie Booth⟨/link⟩⟨/fact⟩
⟨fact name="children"⟩Euan, Nicholas, Kathryn, Leo⟨/fact⟩
⟨fact name="alma_mater"⟩⟨link⟩St John's College, Oxford⟨/link⟩⟨/fact⟩
⟨fact name="profession"⟩⟨link entity_id="E0701822"⟩Lawyer⟨/link⟩⟨/fact⟩
⟨/facts⟩
⟨wiki_text⟩⟨![CDATA[Tony Blair
Anthony Charles Lynton "Tony" Blair (born 6 May 1953) is a British politician who was Prime Minister of the United Kingdom from 2 May 1997 to 27 June 2007. He was Leader of the Labour Party from 1994 to 2007 and the Member of Parliament for Sedgefield from 1983 to 2007...

Figure 2: Excerpt from the KB entry for Tony Blair (E0481157). From LDC2009E58.

2 Related Work

Three types of named entity resolution are found in the literature: *identity resolution,* which matches structured or semi-structured entity descriptions, such as database records; *coreference resolution,* which clusters textual entity descriptions; and *entity linking,* which matches a textual description to a structured or semi-structured description. All three types of resolution have significant literature on monolingual processing; cross-language matching is less well studied.

Identity resolution and its closely related cousin *record linkage* grew out of the database community, which needs to determine when two database records represent the same entity. When matching records are found, identity resolution merges the two records, while record linkage simply notes the correspondence. Brizan and Tansel (2006) present a short overview of work in these fields. Typical approaches combine algorithmic matching of individual column values with hand-coded heuristics to combine the column scores and threshold the result.

Coreference resolution operates over text, determining when two entity mentions refer to the same entity. Approaches to within-document coreference resolution typically exploit syntactic, grammatical and discourse-level features, information that is not available when trying to resolve references across documents. Ng (2010) presents a comprehensive review of recent approaches to within-document coreference resolution. In contrast, cross-document coreference resolution typically assumes that within-document references have been resolved, and tries to place all such mention chains that refer to the same entity into a single cluster that represents that entity. Because the kinds of document-specific features that guide within-document coreference resolution are missing, research in cross-document coreference resolution tends to be more directly applicable to entity linking (which also lacks those features). The Web People Search Evaluation Workshop (Artiles et al., 2010) has been one of the recent drivers of research in cross-document coreference resolution, defining a clustering task that groups Web pages for multiple people bearing the same name.

Entity linking is a hybrid of the preceding two types of named entity resolution, matching a textual entity mention to a set of structured entity representations (usually called the *knowledge base*). Ji and Grishman (2011) present a good overview of the state of the art in monolingual entity linking, as practiced in the TAC evaluation. TAC data sets use a subset of Wikipedia entities for the knowledge base, manually curated query names, and ground truth identified by human assessors without pooling. Wikipedia has been another significant source of training and test data. Adafre and de Rijke (2005) explore automatically adding links between Wikipedia pages (albeit without focusing specifically on named entities). Bunescu and Pasca (2006) trained an SVM to predict whether a query entity matches a Wikipedia page by using hyperlinks within Wikipedia itself as the source of training and test data. Cucerzan (2007) studied identifying entity mentions in text and mapping them to Wikipedia articles. Mihalcea and Csomai (2007) and Milne and Witten (2008) each attempt to identify and properly induce hyperlinks for informative terms in Wikipedia articles (again without specific focus on named entities). Cross-language entity linking has not yet been

widely explored. TAC[1] and NTCIR.[2] have both for the first time announced plans for shared tasks for cross-language entity linking. Steinberger and Pouliquen (2007) describe a system that uses multilingual named entity recognition and cross-language name matching to automatically analyze tens of thousands of news stories daily; however, they do not conduct a formal evaluation of their name merging algorithm.

Contributing to each of these three kinds of named entity resolution are two essential underlying technologies: name matching and context matching. In name matching we ask the question, "Do two different strings represent the same name?" For example, we might like to know whether "Gadhafi" and "Khadafy" are two spellings of the same name. When used as a feature for machine learning, we ask the related question, "How similar are two name strings?"

Cross-language name matching is closely related to name transliteration. Indeed, transliterating a name to the language of the knowledge base, then performing monolingual name matching in that language, is a reasonable approach to cross-language name matching. Name transliteration has an extensive literature; Karimi *et al.* (2011) present a comprehensive survey of the topic.

Name matching does not demand transliteration though; transliteration is a generative process, and name matching requires only that a known name pair be given a score representing the degree of match. Snae (2007) presents a survey of popular name matching algorithms from the record linkage perspective. Monolingually, Levenshtein distance (1966) and its variants are used for basic string matching in many contexts. Cross-language approaches typically combine cross-language mappings of some sort with edit distance metrics. For example, Mani *et al.* (2008) demonstrate a machine learning approach to the problem.

The second crucial underlying technology is context matching. Monolingually, context matching can match on many contextual attributes, including words, entities, topics, or graph structures. Context matching in the translingual setting is closely related to cross-language information retrieval (CLIR); both tasks attempt to estimate the degree of similarity between texts written in different languages. Kishida (2005) presents an overview of the key methods in CLIR.

[1] http://nlp.cs.qc.cuny.edu/kbp/2011/

[2] http://ntcir.nii.ac.jp/CrossLink/

3 Cross-Language Entity Linking

Our approach to entity linking breaks the problem down into two main parts: *candidate identification* and *candidate ranking*. Candidate identification quickly identifies a small set of KB nodes that with high probability contain the correct answer, if it is present. Candidate ranking then considers each candidate in greater detail, producing a ranked list. We give a description of each of these steps in this section; complete details of our English entity linking approach, including descriptions of all of the features used and performance on the TAC-KBP datasets can be found in (McNamee, 2010).

3.1 Candidate Identification

As a KB may contain a large number of entries, we prefer to avoid brute force comparisons between the query and all KB entities. To identify the entries that might reasonably correspond to the input named entity, we rely on a set of fast name matching techniques. We have found that it is possible to achieve high recall without resorting to contextual features. We create indexes for the names in the KB to support fast lookup of potential matches. The specific techniques that we use include:

- Exact match of query and candidate names
- Known alias or nickname lookup
- Number of character 4-grams in common between query and candidate
- Sum of IDF-weighted words in common between query and candidate[3]

In tests on the TAC-KBP 2009 test collection, this approach achieved 97.1% recall. For only 2.9% of the queries, the proper KB referent for the query was not one of the candidates. These cases were particularly challenging because they involved ambiguous organization names or obscure personal nicknames. Our methods are similar to methods used in the database community, sometimes known as *blocking* (Whang et al., 2009) or *canopies* (McCallum et al., 2000).

[3] Inverse document frequency weights enable us to effectively match, for example, Q: Mary Elizabeth Surratt and KB: Mary Surratt, since *Surratt* is a highly discriminating term even though *Mary* is not.

Chinese	306165	Czech	6101
German	34101	Finnish	5639
French	23834	Swedish	5526
Arabic	19347	Danish	2648
Bulgarian	17383	Turkish	2581
Spanish	14406	Macedonian	2469
Italian	12093	Romanian	1981
Dutch	10853	Croatian	1527
Serbian	10020	Urdu	987
Greek	9590	Albanian	257
Portuguese	6335		

Table 1: Number of training pairs for transliterating to English from each language.

To perform cross-language candidate identification, we transliterate[4] the query name to English, then apply our monolingual English heuristics. We used the multilingual transliteration system and training data developed by Irvine *et al.* (2010) in their experiments in orthographic transliteration. The number of training name/transliteration pairs varied by language and is given in Table 1. The source for most of this training data is Wikipedia, which contains links between article pages in multiple languages.

3.2 Candidate Ranking

The second phase in our approach is to score each viable candidate using supervised machine learning, and to select the highest scoring one as output. Each entity linking query is represented by a feature vector $\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{\mathbf{k}}$, and each candidate y is a member of $\mathbb{Y}$, the set of entities in the knowledge base. Individual feature functions, $f_i(\mathbf{x}, y)$, are based on intrinsic properties of the query $\mathbf{x}$, intrinsic properties of a specific KB candidate y, and most commonly, on comparisons between the query and candidate. For each query our goal is to select a single KB entity y or choose NIL if the mentioned entity is not represented in the KB.

Thus, we desire that the correct knowledge base entity y' for a query $\mathbf{x}$ receives a higher score than any other knowledge base entities $y \in \mathbb{Y}, y \neq y'$. We chose a soft maximum margin approach to learning and used the ranking Support Vector Machine approach described by Joachims (2002) and implemented in the SVM$^{\text{rank}}$ tool. We selected a linear kernel for training speed, and set the slack parameter C to be 0.01 times the number of training examples.

In our system absence from the knowledge base is treated as a distinct ranked candidate, the so-called NIL candidate. NIL prediction is integrated into the process by including features that are indicative of no other candidate being correct. Considering absence as a ranked candidate eliminates the need to select a threshold below which NIL will be returned.

The classes of feature functions we use include:

- Name matching features between the query name (Q_{name}) and KB candidate (KB_{name})
- Text comparisons between the query document (Q_{doc}) and the text associated with the KB candidate
- Relation features, chiefly evidence from relations in the KB being evidenced in the Q_{doc}
- Co-occurring entities, detected by running named entity recognition (NER) on the Q_{doc} and finding matching names in the candidate's KB entry
- Features pertaining to the entity type of the KB candidate
- Indications that no candidate is correct and that NIL is therefore the appropriate response

3.2.1 Name matching

A variety of string similarity features are incorporated to account for misspellings, name variants, or partially specified names when trying to match the query name and KB entry. Christen (2006) discusses a variety of name matching features, several of which we adopt. One of the most useful is the Dice score over sets of character bigrams.

3.2.2 Cross-language name equivalence

In all of our cross-language experiments we added name matching features designed to directly calculate the likelihood that a given non-English name is equivalent to a given English name. The model is based on projections of character n-grams across languages (McNamee, 2008).

3.2.3 Contextual Similarity

We measure monolingual document similarity between Q_{doc} and the KB text (KB_{doc}) in two ways: using cosine similarity with TF/IDF weighting; and using the Dice coefficient over bags of words. IDF values are approximated using counts from the Google 5-gram dataset following the method of Klein and Nelson (2008). We also used features such as whether the query string occurs in the KB_{doc} and the KB_{name} occurs in the Q_{doc}.

[4]We use *transliteration* in a broad sense, to include situations where word translation rather than character transliteration is warranted.

To match contexts when the query document and KB are in different languages we treat cross-language context linking as a CLIR problem in which the query is created from the words in the vicinity of mentions of the query name. We adopt Probabilistic Structured Queries (PSQ) (Darwish and Oard, 2003), the key idea of which is to treat alternate translations of a query term as synonyms and to weight the contributions of each "synonym" using a statistical translation model. We index the Wikipedia articles in our test collection using a publicly available IR tool (Indri), learn isolated word translation probabilities from a parallel text using the Berkeley aligner[5] and Joshua,[6] and implement PSQ using Indri's *#wsyn* operator. Based on initial tests on training data, we use a contextual window size of $\pm$ 40 terms to the left and right of the query name mention as the source language query. In Roman alphabet languages, untranslated terms are retained in a character-normalized form.

3.2.4 Relation Features

As can be seen in Figure 2, the KB contains a set of attributes and relations associated with each entity (*e.g.,* age, employer, spouses, etc.). While one could run a relation extractor over the query document and look for relational equivalences, or contradictions, we chose a more straightforward approach: we simply treat the words from all facts as a surrogate "document" and calculate document similarity with the query document.

3.2.5 Named Entity Features

We applied the named entity tagger by Ratinov and Roth (2009) to query documents and created features from the tagger output, including: the percentage of NEs present in KB_{doc}; the percentage of words from all NEs that are present in KB_{doc}; and, the number of co-occurring NEs from Q_{doc} that are present in KB_{doc}. Except for an experiment described in Section 6.1, these features are only used in our monolingual English runs.

3.2.6 Entity Type Features

In English experiments the type of the query entity is determined from the NER output for the query document. Since the reference knowledge base provides a class (*e.g.,* scientist) and a type (*e.g.,* PER) for most entities, we can check whether the type of the KB entity is consistent with the query. This helps discourage selection of eponymous entries named after famous people (*e.g.,* the *USS Abraham Lincoln (CVN-72)*, a nuclear-powered aircraft carrier named after the 16th US president).

3.2.7 NIL Features

Some features can indicate whether it is likely or unlikely that there is a matching KB entry for a query. For example, if many candidates have strong name matches, it is reasonable to believe that one of them is correct. Conversely, if no candidate has high textual similarity with the query, or overlap between KB facts and the query document text, it becomes more plausible to believe that the entity is missing from the KB.

4 Building a Test Collection for Entity Linking in Twenty-One Languages

The TAC-KBP entity linking test collections from 2009 and 2010 include the following resources: (a) a large collection of English documents; (b) approximately 7,000 queries comprising English name mentions from those documents; (c) a reference knowledge base with over 818K entries; and (d) a set of annotations that identify the appropriate KB entry for each query, or absence (McNamee and Dang, 2009; Ji et al., 2010). The KB was created by processing a 2008 dump of English Wikipedia; each entry includes structured attributes obtained from Wikipedia's infoboxes in addition to the unstructured article text. A sample KB entry is shown in Figure 2. We use the TAC KB in all of our experiments.

Since the TAC-KBP queries and documents are only available in English, these data are not directly usable for cross-language entity linking. One approach would be to manually translate the TAC documents and queries into each desired language. This would be prohibitively expensive. Instead, we use parallel document collections and crowdsourcing to generate ground truth in other languages. A fundamental insight on which our work is based is that if we build an entity linking test collection using the English half of a parallel text collection, we can make use of readily available annotators and tools developed specifically for English, then project the English results onto the other language. Thus, we apply English NER to find person names in text (Ratinov and Roth, 2009), our English entity linking system to identify candidate entity IDs, and English annotators on Amazon's Mechanical Turk to select the correct

[5] http://code.google.com/p/berkeleyaligner/

[6] http://sourceforge.net/projects/joshua/

Language	Collection	Queries	Non-NIL
Albanian (sq)	SETimes	4,190	2,274
Arabic (ar)	LDC2004T18	2,829	661
Bulgarian (bg)	SETimes	3,737	2,068
Chinese (zh)	LDC2005T10	1,958	956
Croatian (hr)	SETimes	4,139	2,257
Czech (cs)	ProjSynd	1,044	722
Danish (da)	Europarl	2,105	1,096
Dutch (nl)	Europarl	2,131	1,087
Finnish (fi)	Europarl	2,038	1,049
French (fr)	ProjSynd	885	657
German (de)	ProjSynd	1,086	769
Greek (el)	SETimes	3,890	2,129
Italian (it)	Europarl	2,135	1,087
Macedonian (mk)	SETimes	3,573	1,956
Portuguese (pt)	Europarl	2,119	1,096
Romanian (ro)	SETimes	4,355	2,368
Serbian (sr)	SETimes	3,943	2,156
Spanish (es)	ProjSynd	1,028	743
Swedish (sv)	Europarl	2,153	1,107
Turkish (tr)	SETimes	3,991	2,169
Urdu (ur)	LDC2006E110	1,828	1,093
Total		55,157	29,500

Table 2: Language coverage in our collection.

kbid for each name. Finally, we use standard statistical word alignment techniques implemented in the Berkeley Word Aligner (Haghighi et al., 2009) to map from English name mentions to the corresponding names in the non-English documents.

The six parallel collections we used came from the LDC and online sources. Together, these collections contain 196,717 non-English documents in five different scripts and twenty-one different languages. The final size of the query sets by language is shown in Table 2. We partitioned these queries and their associated documents into three sets per language: 60% for training, 20% for development, and 20% for test. In other work we give additional details about the creation of our test collection (Mayfield et al., 2011).

5 Experimental Results

5.1 English Baselines

Since all of our documents are from parallel corpora, every query is available in English and at least one other language. To serve as a point of comparison, we ran our monolingual entity linking system using the English version of the queries. We also determined performance of a baseline that predicts a *kbid* if its entity's name is a unique, exact match for the English query string, and NIL otherwise.

To compare approaches we calculate the percentage of time that the top-ranked prediction from a system is correct, which we call Precision-at-rank-one (P@1).[7] For the exact match baseline (*Exact*), the mean P@1 accuracy across all query sets is 0.897; on the TAC-KBP 2010 person queries, this baseline achieves a score of 0.832, which is lower most likely because of the intentional efforts at TAC to artificially increase query name ambiguity. Results for both English baselines are included in Table 3.

5.2 Cross-Language Name Matching

Table 3 also reports cross-language experiments in twenty languages where cross-language name matching is used to project the non-English query name into English (*NameMatch*), but the document remains untranslated. If the correct transliteration is known from our transliteration training data, we perform table lookup; otherwise the 1-best transliteration produced by our model is used.

Name matching alone produces serviceable performance. Averaged over all languages, performance on all queries is 93% of the monolingual English baseline. Losses tend to be small in the languages that use a Latin alphabet.

To investigate how errors in automated transliteration affect the system, we also conducted an experiment where the human-produced translations of the entity name were obtained from the English side of the parallel data. In Table 4 we report how this condition (*PerfectTrans*) performs relative to the monolingual baseline. Perfect name translation reduces the error rate dramatically, and performance of 99.2% of monolingual is obtained.

5.3 Name Matching and Context Matching

Table 3 also reports the use of both name matching and context matching using CLIR (+Context). Over all queries, performance rises from 92.9% to 93.9% of the English baseline. Bigger gains are evident on non-NIL queries. In fact, the pan-language average hides the fact that much larger gains are observed for non-NILs in Arabic, Czech, Macedonian, Serbian, and Turkish. We checked whether these gains are significant compared to NameMatch using the sign test; values indicating significant gains ($p < 0.05$) are emboldened.

5.4 Learning Rate

Our approach depends on having a quantity of labelled data on which to train a classifier. To investigate the effect that the number of training ex-

[7] At TAC this metric is called micro-averaged accuracy.

	All Queries					Non-NIL Queries				
		English		Cross-Language			English		Cross-Language	
Set	N	Mono	Exact	NameMatch	+Context	N	Mono	Exact	NameMatch	+Context
ar	577	0.948	0.886 (93%)	0.901 (95%)	**0.926 (98%)**	136	0.838	0.552 (66%)	0.706 (84%)	**0.787 (94%)**
bg	770	0.982	0.918 (94%)	0.892 (91%)	0.892 (91%)	430	0.972	0.854 (88%)	0.821 (84%)	0.833 (86%)
cs	203	0.931	0.764 (82%)	0.828 (89%)	**0.862 (93%)**	136	0.985	0.669 (68%)	0.772 (78%)	**0.838 (85%)**
da	428	0.988	0.963 (97%)	0.965 (98%)	0.963 (97%)	225	0.982	0.933 (95%)	0.938 (95%)	0.933 (95%)
de	217	0.931	0.756 (81%)	0.871 (94%)	0.876 (94%)	154	0.987	0.675 (68%)	0.857 (87%)	0.877 (89%)
el	776	0.979	0.928 (95%)	0.833 (85%)	**0.851 (87%)**	423	0.972	0.868 (89%)	0.714 (73%)	**0.745 (77%)**
es	208	0.909	0.760 (84%)	0.889 (98%)	0.894 (98%)	149	0.960	0.685 (71%)	0.873 (91%)	**0.899 (94%)**
fi	425	0.986	0.965 (98%)	0.927 (94%)	0.941 (95%)	220	0.982	0.936 (95%)	0.868 (88%)	**0.900 (92%)**
fr	186	0.930	0.742 (80%)	0.909 (98%)	0.876 (94%)	135	0.978	0.659 (67%)	0.904 (92%)	0.911 (93%)
hr	846	0.980	0.924 (94%)	0.930 (95%)	0.920 (94%)	470	0.972	0.864 (89%)	0.889 (91%)	0.866 (89%)
it	443	0.984	0.966 (98%)	0.907 (92%)	0.914 (93%)	227	0.978	0.938 (96%)	0.833 (85%)	0.859 (88%)
mk	720	0.978	0.932 (95%)	0.822 (84%)	**0.850 (87%)**	391	0.967	0.875 (90%)	0.706 (73%)	**0.749 (78%)**
nl	441	0.984	0.964 (98%)	0.955 (97%)	0.955 (97%)	224	0.978	0.933 (95%)	0.924 (95%)	0.933 (95%)
pt	443	0.987	0.964 (98%)	0.982 (100%)	0.977 (99%)	230	0.978	0.935 (96%)	0.974 (100%)	0.961 (98%)
ro	878	0.976	0.924 (95%)	0.961 (98%)	0.961 (98%)	480	0.967	0.860 (89%)	0.935 (97%)	0.933 (97%)
sq	849	0.972	0.927 (95%)	0.889 (92%)	**0.913 (94%)**	465	0.955	0.867 (91%)	0.809 (85%)	**0.860 (90%)**
sr	799	0.976	0.920 (94%)	0.804 (82%)	**0.840 (86%)**	447	0.966	0.857 (89%)	0.653 (68%)	**0.743 (77%)**
sv	448	0.987	0.964 (98%)	0.958 (97%)	0.960 (97%)	231	0.978	0.935 (96%)	0.935 (96%)	0.944 (96%)
tr	804	0.980	0.923 (94%)	0.954 (97%)	0.968 (99%)	440	0.973	0.859 (88%)	0.925 (95%)	**0.953 (98%)**
ur	363	0.973	0.862 (89%)	0.810 (83%)	**0.840 (86%)**	215	0.967	0.772 (80%)	0.707 (73%)	**0.763 (79%)**
$\bar{x}$	541	0.968	0.897 (93%)	0.899 (93%)	0.909 (94%)	291	0.967	0.826 (85%)	0.837 (87%)	0.864 (89%)

Table 3: P@1 for a variety of experimental conditions. The left half of the table presents aggregate results for all queries; on the right performance is given for just non-NIL queries. Percentages are with respect to the monolingual English condition.

Set	Mono	NM	PerfectTrans		English NEs	
	P@1	P@1	P@1	% Mono	P@1	% Mono
ar	0.948	0.901	**0.941**	99.3%	0.901	95.1%
bg	0.982	0.892	**0.986**	100.4%	**0.855**[8]	87.0%
cs	0.931	0.828	0.882	94.7%	**0.897**	96.3%
da	0.988	0.965	**0.986**	99.8%	0.972	98.4%
de	0.931	0.871	0.899	96.5%	0.917	98.5%
el	0.979	0.833	**0.978**	99.9%	**0.872**	89.1%
es	0.909	0.889	0.914	100.5%	0.861	94.7%
fi	0.986	0.927	**0.988**	100.2%	**0.955**	96.9%
fr	0.930	0.909	0.909	97.7%	0.914	98.3%
hr	0.980	0.930	**0.972**	99.2%	**0.963**	98.3%
it	0.984	0.907	**0.987**	100.2%	**0.930**	94.5%
mk	0.978	0.822	**0.976**	99.9%	**0.881**	90.1%
nl	0.984	0.955	**0.982**	99.8%	0.964	97.9%
pt	0.987	0.982	0.987	100.0%	0.977	99.1%
ro	0.976	0.961	**0.976**	100.0%	0.960	98.4%
sq	0.972	0.889	**0.976**	100.5%	**0.933**	96.0%
sr	0.976	0.804	**0.974**	99.7%	**0.977**	100.1%
sv	0.987	0.958	**0.984**	99.8%	**0.975**	98.9%
tr	0.980	0.954	**0.984**	100.4%	0.963	98.2%
ur	0.973	0.810	**0.931**	95.7%	**0.876**	90.1%
$\bar{x}$	0.968	0.899	0.961	99.2%	0.927	95.8%

Table 4: Cross-language effectiveness (P@1) over all queries with optimal (a) transliteration and (b) named entity recognition. Simply having access to perfect transliterations achieves 99% of monolingual performance, on average. Providing lists of named entities mentioned in the document, in English, also improves performance. Bold values indicate statistically significant gains compared to the NameMatch (NM) run.

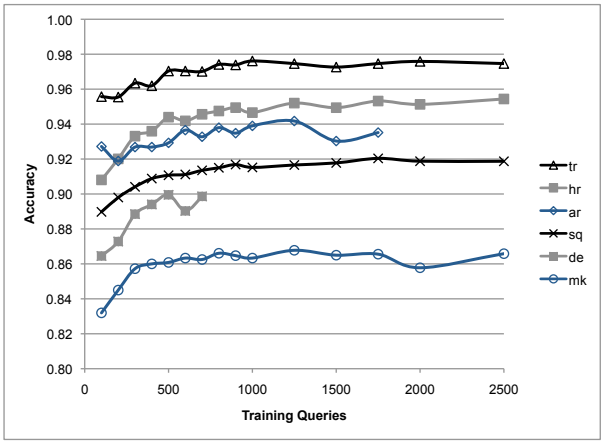

Figure 3: Classifier accuracy and training set size.

emplars has on classifier accuracy we built classifiers using fixed numbers of training queries. Figure 3 shows these results for selected languages. Each curve was produced by generating a random permutation of the training data, selecting the first k queries, and averaging the results over five trials. Note that the total amount of available training data differs by language. In all cases, accuracy rises quickly for the first 500 queries, and little improvement is observed after 1000 examples.

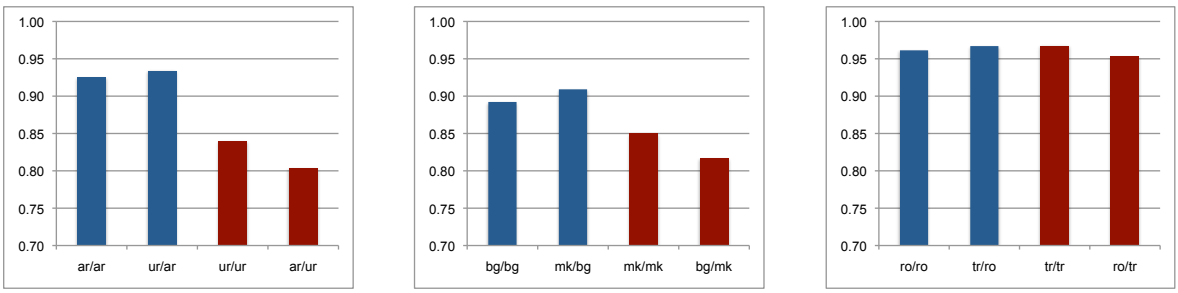

Figure 4: Training with annotations from another language using the same writing system. A label of xx/yy indicates that feature weights were trained using labelled data from language xx and then applied to queries on language yy.

6 Additional Experiments

6.1 Multilingual NER and Transliteration

We believe that the entities in a document that co-occur with a target entity are important clues for disambiguating entities. However, while we had ready access to named entity recognizers in English, we did not have NER capability in all of the languages of our collection. We would like to know how useful multilingual NER could be. To simulate this using our test collection, where all documents are from parallel texts with an English translation, we conducted an experiment that used the English documents only to recognize English named entities that co-occur with the query string; in every other respect, the untranslated foreign document was used by the system. The English NER may make errors, but we use it to simulate the performance of excellent non-English NER coupled with perfect entity translation.

Table 4 shows that co-occurring entities are a very helpful feature. Compared to name matching alone, average P@1 rises from 89.9% to 92.7%.

6.2 Cross-Language Training

Although we have demonstrated an efficient method for building a test collection for cross-language entity linking, it may still be difficult to obtain training data and tools for some less-resourced languages. Our process and feature set is largely language-independent, and we would like to know how feasible it is to make predictions without any language-specific training data by exploiting entity linking annotations from a related language. We examined pairs of languages using the same script – Arabic/Urdu, Bulgarian/Macedonian, and Romanian/Turkish – and trained classifiers using labeled data for the other language. Figure 4 shows that performance is not dramatically different when using annotations from a language sharing a common alphabet. This suggests that it is plausible to build a cross-language entity linking system without manually-produced annotations for a particular language.

7 Conclusions

In this paper we introduced a new problem, cross-language entity linking, and we described an approach to this problem that uses statistical transliteration and cross-language information retrieval. Using a newly-developed test collection for this task,[9] we demonstrated the success of the approach in twenty languages. Our best model using both name and context matching achieves average performance across twenty languages which is 94% of a strong monolingual English baseline, with individual languages ranging from 86% to 99%. Additionally, we characterized the number of training exemplars needed, demonstrated the feasibility of off-language training, and illustrated performance gains that are possible if combined multilingual NER/transliteration is available.

Acknowledgments

We are grateful to Chris Callison-Burch and Ann Irvine for their support with machine translation and orthographic transliteration, and to Tan Xu and Mohammad S. Raunak for their help in data curation. Support for one of authors was provided in part by NSF grant CCF 0916081.

[8] The English NEs run was significantly *worse* vs. NameMatch in Bulgarian (bg).

[9] The test collection is available at `http://web.jhu.edu/HLTCOE/datasets.html`.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in Wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97. ACM.

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigo. 2010. Overview of the web people search clustering and attribute extraction tasks. In *CLEF Third WEPS Evaluation Workshop*.

David Guy Brizan and Abdullah Uz Tansel. 2006. A survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 6(3):41–50.

Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, Australian National University.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Empirical Methods in Natural Language Processing*.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *ACM SIGIR*, pages 338–344. ACM.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 923–931. ACL.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *AMTA*.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Association for Computational Linguistics*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 Knowledge Base Population track. In *Text Analysis Conference (TAC)*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Knowledge Discovery and Data Mining (KDD)*.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys*, 43(4):1–57.

Kazuaki Kishida. 2005. Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41(3):433 – 455. Cross-Language Information Retrieval.

Martin Klein and Michael L. Nelson. 2008. A comparison of techniques for estimating IDF values to generate lexical signatures for the web. In *WIDM '08*, pages 39–46. ACM.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics–Doklady*, 10(8):707–710.

Inderjeet Mani, Alex Yeh, and Sherri Condon. 2008. Learning to match names across languages. In *MMIES '08*, pages 2–9. ACL.

James Mayfield, Dawn Lawrie, Paul McNamee, and Douglas W. Oard. 2011. Building a cross-language entity linking collection in twenty-one languages. In *Cross-Language Evaluation Forum (CLEF)*.

Andrew McCallum, Kamal Nigam, and Lyle Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Knowledge Discovery and Data Mining (KDD)*.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the TAC 2009 Knowledge Base Population track. In *Text Analysis Conference (TAC)*.

Paul McNamee. 2008. *Textual Representations for Corpus-Based Bilingual Retrieval*. Ph.D. thesis, University of Maryland Baltimore County, Baltimore, MD.

Paul McNamee. 2010. HLTCOE efforts in entity linking at TAC KBP 2010. In *Text Analysis Conference (TAC)*, Gaithersburg, Maryland, November.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242.

David N. Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *CIKM*, pages 509–518.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Chakkrit Snae. 2007. A comparison and analysis of name matching algorithms. *Proceedings of World Academy of Science, Engineering and Technology*, 21:252–257, January.

Ralf Steinberger and Bruno Pouliquen. 2007. Cross-lingual named entity recognition. *Linguisticae Investigationes*, 30(1):135–162, January.

Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. 2009. Entity resolution with iterative blocking. In *SIGMOD 2009*, pages 219–232. ACM.