Cross-Language Person-Entity Linking from Twenty Languages

DAWN LAWRIE*
Computer Science Department, Loyola University Maryland
Email: lawrie@cs.loyola.edu

JAMES MAYFIELD
Johns Hopkins University Human Language Technology Center of Excellence
Email: james.mayfield@jhuapl.edu

PAUL MCNAMEE
Johns Hopkins University Human Language Technology Center of Excellence
Email: paul.mcnamee@jhuapl.edu

DOUGLAS W. OARD
College of Information Studies and UMIACS, University of Maryland, College Park
Email: oard@umd.edu

*This work has been supported in part by NSF grant CCF 0916081

Abstract

The goal of entity linking is to associate references to some entity that are found in unstructured natural language content to an authoritative inventory of known entities. This paper describes the construction of six test collections for cross-language person-entity linking that together span 22 languages. Fully automated components were used together with two crowdsourced validation stages to affordably generate ground truth annotations with an accuracy comparable to that of a completely manual process. The resulting test collections each contain between 642 (Arabic) and 2,361 (Romanian) person references in non-English texts for which the correct resolution in English Wikipedia is known, plus a similar number of references for which no correct resolution into English Wikipedia is believed to exist. Fully automated cross-language person-name linking experiments with 20 non-English languages yielded a resolution accuracy of between 0.84 (Serbian) and 0.98 (Romanian), which compares favorably with previously reported cross-language entity linking results for Spanish.

Keywords: test collections, entity linking, multilingual processing

Introduction

The entity linking task requires that we identify which one of a set of known entities is the referent of some mention in running text. In essence, this is a fully automated counterpart to the well known problem of name authority control [41]. In general the entity may be of any type and the mention may be in any form, but in this paper we restrict our attention to named references (excluding nominal and pronominal references) and to entities that are people (thus, for example, excluding organizations). Because Wikipedia has been used extensively as a target "knowledge base" in entity resolution evaluations, we use as our target the English pages that describe people in the 2008 Wikipedia snapshot (LDC2009E58), which has been used in the Text Analysis Conference (TAC) Knowledge Base Population (KBP) task since 2009 [28].

The focus of our research is on cross-language entity resolution, in which the Wikipedia pages for the entity are in English, but the running text from which we wish to link a reference are in some other language. In particular, we are interested in the degree to which techniques generalize well across languages. To evaluate these techniques, we need cross-language entity linking test collections for many languages. We therefore focus in this paper on an affordable technique for building multilingual entity linking test collections from parallel document collections, refining the methods that we have previously described in [24]. We have used this technique to build six test collections for cross-language entity linking that together span 22 languages. To support repeatable evaluation, we have divided these test collections into training, development test, and evaluation sets.

The usual approach to entity linking draws on evidence from both orthography (i.e., how a named reference is written) and context (e.g., what is being said about that entity) [29]. Cross-language entity linking thus requires that we accommodate differences in orthography between languages on the one hand and differences in the way those words in different languages are used to express similar meanings on the other hand. We therefore introduce generalizable techniques for accommodating those two types of differences, drawing on learned statistical transliteration models and learned statistical translation models, respectively.

We have used these techniques with five of the six test collections to look broadly at differences between cross-language person-entity linking accuracy for 20 language pairs. For 12 of the 20 language pairs, we obtained better than 95% of the one-best[1] linking accuracy that a state-of-the-art monolingual person-entity linking system could achieve on the same set of mention-queries (English one-best accuracy for the 20 mention-query sets varied between 0.91 and 0.99). Only one language pair (Serbian-English, at 0.84) yielded one-best cross-language accuracy below 90% of the one-best monolingual English condition. We present some analysis of these results that focuses on specific language characteristics (e.g., character set or morphological richness), but our key conclusion from these experiments is that robust techniques for cross-language entity

[1] A standard measure of effectiveness for entity linking is the fraction of the cases for which the correct resolution is present among the system's top n guesses. When $n = 1$, we refer to this as the "one-best" accuracy.

linking that scale well to multiple language pairs can be affordably built. A second contribution of our work is the set of three multilingual and three bilingual test collections for cross-language person-entity linking that we have built, which are freely available for research use.[2] Our third contribution is the method we have used to create those test collections, which could now be applied to affordably create test collections for other language pairs, other document types, and other entity types.

The remainder of this paper is organized as follows. The next section presents related work on monolingual and cross-language entity lining. Then we explain how we build cross-language test collections by using monolingual tools, parallel document collections, and crowdsourcing, and the following section describes the six test collections that we have built using those methods. The penultimate section then briefly describes the design of our fully automated cross-language entity linking system (which we have previously described in greater detail in [29]) and presents new results from that system for 20 language pairs. Finally, the paper concludes with a few remarks about the implications of this work and a brief discussion of future research that could build on our results.

Related Work

In this section we first review the related work on monolingual entity linking, the case in which the document containing the reference is written in the language for which the knowledge base was originally designed. We then review the more recent work on cross-language entity linking that informs our research. The section concludes with a review of existing entity linking test collections.

Monolingual Entity Linking

Three broad types of named entity resolution are found in the literature: *record linkage*, which seeks to match records that represent entities in structured collections such as database records; *coreference resolution*, which seeks to match references to the same entity in running text; and *entity linking*, which seeks to match a mention of an entity that is found in running text to a record that represents that entity in a structured collection of entity records. All three types of entity resolution problem have substantial (and often disjoint) literatures.

Record linkage has long been a focus of database researchers, who sometimes need to automatically determine when two database records represent the same entity. Database research draws a distinction between record linkage and the subsequent merging of those linked records (which is often referred to as *identity resolution*). Brizan and Tansel [6] present a short overview of work on these topics. Typical approaches in small-scale cases use hand-coded heuristics based on attribute values to compute similarity

[2] http://hltcoe.jhu.edu/publications/data-sets-and-resources

scores, which can then either be examined manually or automatically accepted as identical based on some threshold; larger-scale applications offer greater scope for the application of machine learning. As with many language processing tasks, the combination of several approaches to identity resolution often outperforms any one approach [8]. Record linkage is also a key component of the more comprehensive *ontology merging* problem that is at the core of research on *Linked Open Data* (LOD), which has developed in recent years as one step towards the vision of a *semantic web* [15].

Coreference resolution, by contrast, has long been the province of research in computational linguistics. As usually formulated, the coreference resolution task follows the automated detection of named, nominal or pronominal references to entities, and it sometimes precedes automated detection of specific types of relationships between those entities (e.g., employed-by or married-to). Computational linguistics research draws a distinction between *within-document* and *cross-document* coreference. Approaches to within-document coreference resolution typically exploit proximity, syntax, and discourse features to construct coreference chains in which mentions of the same entity are clustered. Ng [34] presents a comprehensive review of recent approaches to within-document coreference resolution. Cross-document coreference resolution methods typically take within-document coreference chains as input, seeking matches in both the way a reference to an entity is written (*orthographic evidence*) and in what is said about that entity (*contextual evidence*). The Web People Search (WePS) evaluation workshop [3] has been one recent driver of research in cross-document coreference resolution, defining a clustering task in which the goal is to partition a set of Web pages that refer to people by the same name into sets of pages that refer to each different person. The ACE 2008 workshop also conducted evaluations of cross-document entity coreference resolution for both Arabic and English [5].

Entity linking research has emerged more recently as a component task in the automatic construction of knowledge representations from text, a problem referred to generally as *Knowledge Base Population* (KBP). As usually formulated, some structured collection (the *knowledge base*) already exists, and the goal of entity linking is to detect whether some reference to an entity in a document refers to any known entity, and if so which one. Entity linking is thus an asymmetric hybrid between cross-document coreference (typically drawing on evidence from within-document coreference chains in the *source* document) and record linkage (typically drawing on attributes and relations that are already coded in the *target* knowledge base). To date, the entity linking task has garnered attention principally from computational linguistics and Web science researchers. In 2009, the Text Analysis Conference (TAC) Knowledge Base Population (KBP) track conducted an evaluation of English entity linking using a set of news stories as the sources and a snapshot of English Wikipedia pages that represent named entities as the target knowledge base, drawing on the *infobox* for structured content and on the content of the Wikipedia page as a source for associated text [28]. Ji and Grishman [18] present a good overview of the state of the art in monolingual entity linking, as practiced in TAC.

Entity linking is also, in one sense, a restricted version of the *wikification* task that was introduced in the INEX evaluations, which principally attracts information retrieval researchers. In wikification the goal is to predict which links a Wikipedia editor will choose to make, given the text of the page (with no links) and the opportunity to train on a large number of other Wikipedia pages (that do include links). In entity linking the evaluation framework often assumes that the mention to be linked has already been somehow identified, whereas in wikification knowing what to link is a part of the challenge. Among the extensive research on wikification, Adafre and de Rijke [1] reported some of the first work on this topic, and Milne and Witten [33] published one of the most highly cited papers for that task (notably, also using the same technique to link from news stories to Wikipedia).

Cross-language Entity Linking

Each task described above has a cross-language counterpart. Fundamentally, each task involves two underlying technologies: name matching and context matching. In name matching we ask the question, "Do two strings represent the same name?" For example, we might like to know whether "Gadhafi" and "Khadafy" are two spellings of the same name. And when used as a feature for machine learning, we ask the related question, "How similar are two name strings?" Context matching attempts to measure the degree of match between the context surrounding the name mention to be linked and the context surrounding the knowledge base items that are candidate link targets. In text, the context might be nearby words; in a database, the context might be other field values in the same record. Cross-language entity linking draws on the same two fundamental capabilities, but with the added challenge that the similarity metrics must operate between languages.

The greatest challenge for cross-language name matching arises when trying to match across languages that are written using different character sets. One straightforward approach in such cases is to use transliteration to rewrite names from one character set to another and then to perform same-language name matching on the result. Name transliteration has an extensive literature; Karimi *et al.* [21] present a comprehensive survey of the topic. Most transliteration work has focused on specific language pairs, but statistical techniques that are reasonably language independent can also be crafted [17]. Transliteration is a generative process, but name matching requires only that a name pair be given a score representing the degree of match. Such an approach is sometimes referred to as *cognate matching* [37].

Another challenge is that names in some languages can be written differently depending on the way they are used in a sentence. This morphological variation can be accommodated using language-specific morphological analysis techniques, or alternatively, with simpler techniques that look for consecutive sequences of identical characters (i.e., character n-grams). Snae [38] presents a survey of popular name matching algorithms from the record linkage perspective. Monolingually, Levenshtein distance [25] 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.* [26] demonstrated a machine learning approach to the problem.

Context matching might draw on many contextual attributes, such as words, entities, topics, or graph structures. Where context matching is performed on numerical, categorical, or structural features, as is sometimes the case in record linkage, language differences may be of little consequence. However, when the context of interest is nearby words, as is the case with written text, then cross-language context matching requires some way of matching words that have similar meanings across languages. Translate-then-match offers one option when a full machine translation system is available, but again simpler alternatives are also possible. One simple approach is to draw on an extensive body of work on Cross-Language Information Retrieval (CLIR) that Kishida [22] and Nie [35] have surveyed to define a cross-language similarity measure between spans of text. Several techniques are available for this, but one effective approach is to learn word substitution probabilities from existing translations and then to embed those probabilities in the matching function [44].

As with their monolingual counterparts, cross-language record linkage, co-reference resolution, and entity linking have all been the subject of study. Notable examples of cross-language record linkage include the MITRE Challenge [32] (which focused on name matching) and the Ontology Alignment Evaluation Initiative (OAEI), which developed a test collection (MultiFarm) that calls for both name matching and context matching [31]. A precursor to evaluations focused specifically on cross-language entity linking was the WebCLEF task held in 2005 and 2006 [4], which focused somewhat more broadly on known-item search, such as finding a named Web page. For example, given the query "El Palacio de la Moncloa" (Moncloa Palace), a system should return the URL: `http://www.lamoncloa.gob.es/`. Research on cross-language entity linking has been a focus of TAC annually since 2011 (in a knowledge base population task), and of another evaluation known as CrossLink in 2011 and 2013 (in a wikification task) [40].

Development of Test Collections

Once it becomes possible to build systems to perform a task, the question then naturally arises is how well those systems actually work. We need to know this for two reasons. Most obviously, if two or more approaches are available, we want to know which one we should use. We call this *summative evaluation*. Perhaps even more importantly, when we are building a new system we would like to tune it to do as well as it can—for this we need *formative evaluation*. While summative evaluation might be done just once, formative evaluation typically needs to be done repeatedly. Thus, for formative evaluation we need some approach to evaluation that is affordable, and experience has shown that one useful way of making evaluation more affordable is to build a reusable *test collection* that makes it possible to amortize the cost of developing

evaluation resources over many uses. One of the earliest test collections, specialized to evaluation of what we would today call a search engine, was developed for the Cranfield I [9] and Cranfield II [10] experiments.

A second important innovation was the *shared task* in which many researchers use the same test collection to generate results that are broadly comparable. Shared tasks produce three outcomes that together have come to be seen as characteristic of an *evaluation-guided research paradigm*: (1) a community of researchers who come together around the task; (2) an agreed way of evaluating progress on the task; and (3) baseline results to which future progress can be compared. Notably, test collections stand at the center of this process as the most visible artifact around which evaluation-guided research is built. Test collections for entity linking trace their heritage back to the Message Understanding Conferences (MUC), which were held in the 1990's to (among other things) assess the ability of automated systems to detect mentions of named entities in text [16].

With the shift from symbolic to statistical techniques, it became important that test collections be sufficiently large to adequately represent even relatively rare phenomena. This need was driven in part by a desire for increasingly realistic evaluation, but the larger driving force was a need for an adequate quantity of annotations on which to train machine learning techniques. Where hundreds of annotated examples might suffice for evaluation, thousands of examples are typically required as training examples. Moreover, as learning techniques became more sophisticated, the need for a third set to support development testing (*devtest*) often arose as well. As test collections became larger, reuse across multiple research groups became particularly desirable as a way of further amortizing development costs. This led to the creation of organizations devoted to corpus production, annotation, and dissemination. For example, the Linguistic Data Consortium (LDC)[3] now produces approximately thirty annotated collections every year. Cole [11] presents a good overview of corpus collection, and Glenn et al. [13] present a typical LDC process for collection creation and annotation. Ellis *et al.* [12] describe the processes that NIST and LDC use to create entity linking collections. The Evaluation and Language resources Distribution Agency (ELDA)[4] is a similar organization, focusing more strongly on the languages of the European Union.

Construction, annotation and curation of a high quality test collection typically follows a meticulous process. Although widely shared evaluation resources are valuable, their development can only be justified when there is reason to believe that a research community that would make good use of them exists or is likely to form around them. The evaluation-guided research paradigm has for this reason (and others) been criticized for a tendency towards incrementalism.; focusing extensive effort on those problems for which good evaluation resources exist, at the cost, perhaps, of failing to explore new questions that would require new resources. In recent years, this consideration has led to an interest in finding ways to build test collections

[3]http://www.ldc.upenn.edu/
[4]http://www.elda.org/

more affordably, one large thrust of which has been crowdsourcing. Amazon's Mechanical Turk[5] is probably the best known crowdsourcing platform; others include Clickworker,[6] Cloudcrowd,[7] Microtask,[8] Samasource[9] and Mobileworks.[10]

Snow et al. [39] were among the first to demonstrate the usefulness of crowdsourcing on a variety of language processing tasks; the use of crowdsourcing for collection curation and annotation has blossomed since. Quality control is perhaps the biggest challenge in putting together a new annotated corpus, with accuracy and consistency being the two principal dimensions to be controlled. Medero *et al.* [30] break quality control into a sequence of small decisions; Snow *et al.* [39] stress the need for validation of crowdsourced annotations and propose a weighted voting method that outperforms the basic one-annotator-one-vote methodology.

Existing cross-language entity linking test collections, developed for the cross-langauge entity linking evaluations described above, currently exist for just a few language pairs. Our goal in this paper is to dramatically extend the range of language pairs for whcih cross-language entity linking can be evaluated under comparable conditions.

Building Test Collections

We define the entity linking task formally as:

> Given one or more mentions of an entity in a document (a *query*) and a set of known entities (a *knowledge base*), find the entity ID of the mentioned entity within the knowledge base (KB), or return NIL if the mentioned entity was previously unknown.

Our approach to collection creation has three distinguishing characteristics: (1) the use of parallel document collections to allow most of the work to occur in a single language, (2) the use of crowdsourcing to quickly and economically generate many human judgments, and (3) controlling costs using triage. A fundamental insight behind our work is that if an entity linking test collection using the English half of a parallel document collection is built, we can make use of readily available tools developed specifically for English and then project the English results onto the other language. Thus, English named entity recognition (NER) is applied to find person names in text, an English entity linking system identifies candidate entity IDs, and English annotators available through crowd-sourcing select the correct entity ID for each name. Standard statistical word alignment techniques are then used to map name mentions in English documents to the

[5] https://www.mturk.com/
[6] http://clickworker.com/
[7] http://cloudcrowd.com/
[8] http://microtask.com/
[9] http://samasource.org/
[10] https://www.mobileworks.com/

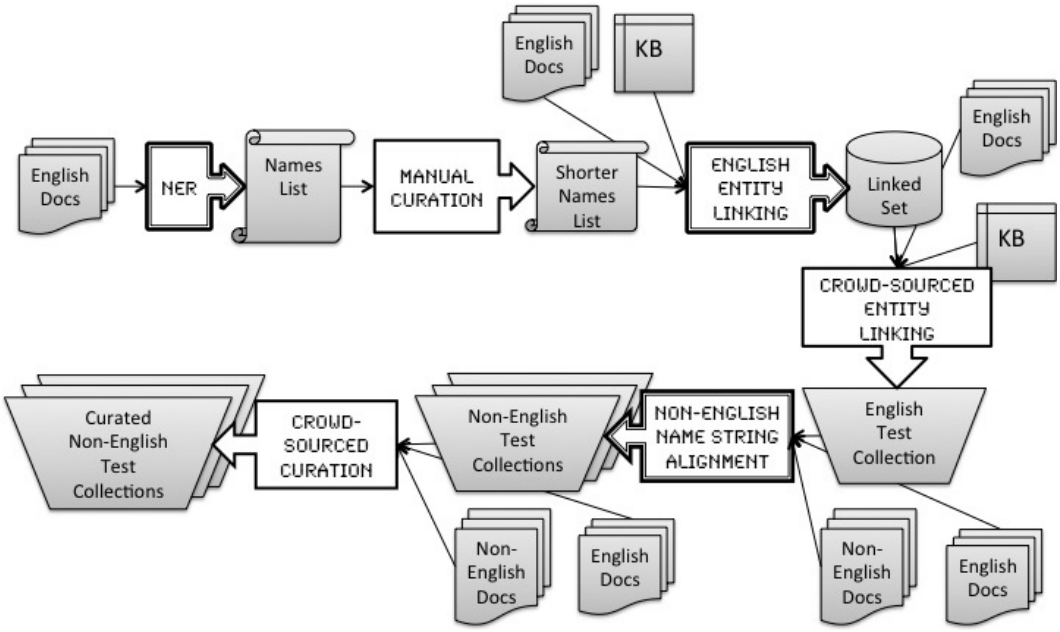

Figure 1: Overview of Test Collection Creation Process.

corresponding names in non-English documents. Projection of named entity annotations is known to be imperfect [45]; therefore, crowd-sourcing is applied again to curate the name projections. To control costs, we curate only the putatively positive instances at every stage; putatively negative instances are simply removed. Because this decision raises the possibility that the test collection might become unbalanced (e.g., favoring "easier" cases), we return to this question in the sections concerning collection statistics and generating ground truth links by looking in detail at the triage decisions made in our approach in the process of constructing six actual test collections. An overview diagram of the process of test collection creation appears in Figure 1. The inputs and outputs are shaded, while the processes appear in arrowed boxes. Those arrowed boxes that have double borders represented automated steps in the process. Others steps are manual, which combine crowd-sourced labor with our own.

Generating English Queries

Following the formulation of the TAC Entity Linking task, a mention-query (henceforth, simply a "query") is a {name, document} pair in which the name string occurs at least once (and possibly more often than that) somewhere in the document. In choosing the named references to be resolved for our test collection, our goal was to select a substantial number of naturally occurring full-name references that are representative

of those that might need to be resolved. All of our test collections include English as one of the languages, so we started by applying the publicly available named entity recognition (NER) system created by Ratinov and Roth [36] to automatically tag all named mentions of a person entity. This system attained the highest published score on the CoNLL 2003 dataset [42], and thus seemed to us to be a reasonable choice. For each document, we formed a list of each unique name mention string.

For person entities, single-token name mentions rarely need to be resolved by operational systems because within-document co-reference resolution can typically be used to identify a more complete named mention within the same document. As a first triage step, we therefore removed all single-token names that had been detected by NER. We then removed all names that occurred in only one document because we wish to produce references to entities that are sufficiently prominent that they might be represented in a knowledge base. Finally, we removed any named mention for which the character string was a proper substring of the character string for some other detected named mention of a person in the same document, thus retaining only the longest mention.

The resulting putative name strings from NER contain two types of errors: some mentions will have been missed, and some putative mentions will not actually be person names. We consistently (here, and throughout our test collection development process) simply do our best to minimize misses and then accept the consequences of those misses that do occur. To limit the introduction of noise from incorrectly recognized named references, an author of this paper manually examined a deduplicated list of the remaining extracted (English) names to eliminate strings that were obviously not person names. Because this was done on a deduplicated list by a native speaker of English, this triage review was quite fast.

The distribution of named references is typically sharply skewed, with a few people each being mentioned in many documents and many people each being mentioned only in a few documents. Because we would like the test collection to consist of names from both categories (the popular and the obscure), we chose to limit the number of duplicates of a specific name string as a query to ten. Thus, a popular entity will be included in no more than ten queries, each of which will reference a different document. When we report average effectiveness measures over the full query set, this limitation has the effect of giving greater emphasis to relatively rare entities than simple random selection would have yielded. The result of this process is a set of queries (i.e., {name, document} pairs) that could be used to evaluate an entity linking process. The next step is to establish a ground truth resolution for each such query.

Generating Ground Truth Links

Optimally, we might wish to construct ground truth links between queries and person-entities in some knowledge base that are representative of what an expert annotator would construct if given unlimited time, capable search tools, and rich sources of evidence. Our goal in this work is somewhat more limited, however,

because our principal interest is in measuring the accuracy of cross-language entity linking systems, and for that more limited evaluation task, we can usefully express our results by comparing the accuracy of a cross-language entity linking system with that of a comparable English entity linking system. We can, therefore, reasonably start with the entities that an English entity linking system can find, and doing so considerably reduces annotation costs.

We have chosen the TAC KBP knowledge base as our target, both because we have an existing English entity linking system for that target and because the TAC KBP knowledge base has been used in other cross-language entity linking experiments to which our results can be compared. The TAC KBP knowledge base is derived from an October 2008 subset of Wikipedia pages that contained infoboxes; more than 114,000 persons are represented.

We begin by generating a ranked list of candidate entities using the our HLTCOE English entity linking system [27]. Our system yielded competitive results in the TAC 2009, 2010, and 2011 entity linking evaluations, with a Recall@3 on person entities for which a match could be found by NIST assessors of 94.4% (201/213) in TAC 2010. Increasing the depth beyond 3 would have increased the measured recall on person entities to 94.8% (202/213), so a cutoff of 3 achieves nearly the upper bound of coverage for this English entity linking system; the remaining 11 correct answers were missing entirely from the entity linker's ranked output. Given these results, we chose to consider only the top three results from our HLTCOE entity linking system as potential candidates when building our test collections.

The conditions evaluated in TAC 2010 were somewhat different from those of our collections, and there are two reasons to believe that the Recall@3 in our collections is somewhat higher than 94.4% First, approximately half of our recall failures on the TAC 2010 entity linking task were single-token names. By design, we have no single token-names among our queries. Second, the TAC 2010 queries were designed with an intentional bias in favor of confusable names [19]. We, by contrast, seek to reflect the naturally occurring distribution of query difficulty by using all named references that we can as queries. Considering these two factors, we conservatively estimate that the recall failure rate of 5.6% for our English entity linking system on the TAC 2010 collection overstates the recall failure rate on the queries in our new dataset by a factor of two.

Our next task was to select among the top three candidates. Human judgments were collected for this purpose using Amazon's Mechanical Turk [2], which has been applied to a wide array of HLT problems [7, 39]. The Human Intelligence Task (HIT) that we constructed for this purpose asked the assessor to read the text surrounding a named mention of the entity in the document and to read up to 2,000 characters from the beginning of the Wikipedia entry for each of the top three person entities returned by our English entity linking system. Our English entity linking system can guess NIL (meaning that no entry is believed to exist in the target knowledge base) at any position in the ranked list, but for this purpose we ignored NIL results

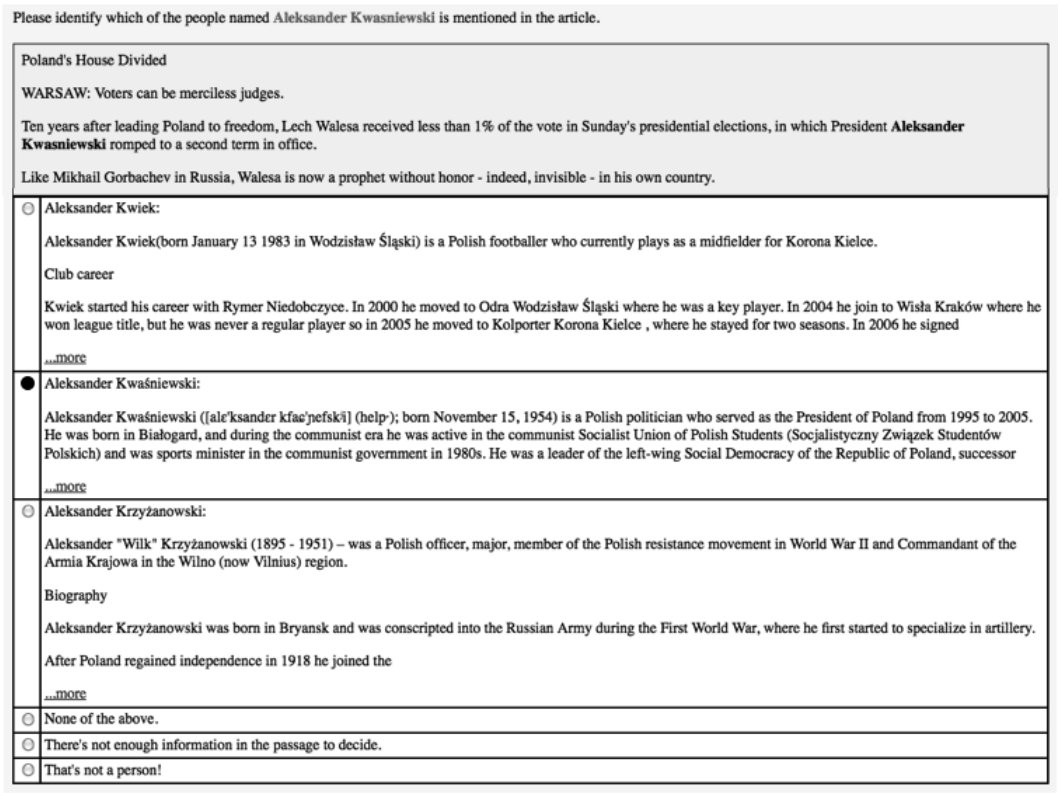

Figure 2: Mechanical Turk entity linking task.

and used only the top three non-NIL results. An example of the user interface for this HIT is shown in Figure 2. To minimize system-induced bias, the three candidate entities were not presented in the best-first order returned by our entity linking system, but rather in a randomized order. In addition to the three candidate entities, the assessor could select "None of the above" (in which case we would record NIL as the ground truth link) or that the query was problematic by selecting "There's not enough information" or "That's not a person" (in which cases we would remove the query from the test collection). A single Mechanical Turk HIT consisted of six such tasks, two of which were interleaved queries for which the correct response was already known. These known correct responses were used for quality control.

Because of task size limitations on Mechanical Turk, the English HITs for a collection were divided into batches, with typical batch sizes of around 600 HITs. We computed accuracy scores for each Turker in each batch based on the fraction of correct choices that a Turker made on the two queries per HIT for which the correct resolution was already known, and Turkers with an accuracy below 80% were eliminated. Three

independent judgments for each query were obtained, and a query was included in the collection only if none of the three Turkers had been eliminated for low accuracy and only if all three Turkers agreed on the answer.

Generating Non-English Queries

Once we have an entity linking test collection for English, all that remains to be done is to identify the corresponding named mentions in the non-English documents. There are several options available. For example in the cases where the English query had been linked to the knowledge base, the curated Wikipedia cross-language links could be used. This option was not utilized for two reasons. One is because it could only be used for about half the queries where the links exist. The second reason is this could introduce bias into the collection because the query string would be more likely to have an exact match with the Wikipedia entry.

Instead we chose a process that could be utilized for all queries, which relies on the parallel alignment of the document collections. Our first step in that process would normally be to perform sentence alignments for each document-aligned English/non-English document pair, but we did not need to perform that step because sentence alignments were provided by the distributors of all six of the parallel document collections that we used. We tokenized all of the sentences in every language, which we did simply by splitting on white space for every language except Chinese and by using the Stanford segmenter for Chinese. After alignment and tokenization, we used the Berkeley Word Aligner [14] to create a mapping from single tokens in the English text to tokens or sequences of tokens in the other language.

For each named mention identified by the NER system, whether used in a query or not, we then marked the shortest contiguous span of tokens in the target language document that included every token or token sequence that aligned with a token in the English named mention. If the alignment was perfect and if all names were written contiguously in the target language, this approach could recover some components of a named reference (e.g., a middle name) that had been omitted on one side or the other. If the alignment were imperfect, very long erroneous alignments could result. Such cases are, however, easily spotted.

By aligning all names, rather than only those in the query set, the entire parallel document collection can be used to compensate for a misalignment in the query document. Our final step, therefore, ranks all the projections for each English name string in decreasing order of frequency (i.e., the number of occurrences in the entire parallel text for a given language pair) and the most frequent aligned token sequence that actually exists in the aligned non-English document is chosen as the non-English named mention. Ties are broken based on the difference in the number of tokens in each language, with smaller absolute values of the difference in the token count being preferred. Remaining ties are resolved in favor of longer non-English mentions, and ties that still remain are broken arbitrarily. The resulting non-English {name, document} pair is then taken as a query candidate, pending further review.

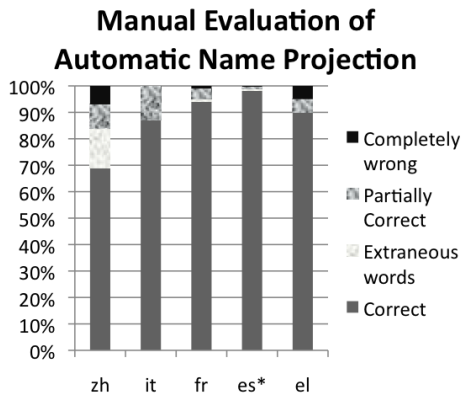

Figure 3: Accuracy of automated name projection as determined by manual inspection. * indicates that all queries were inspected for Spanish. In other cases a random sample of 100 queries was inspected. See Table 2 for language codes.

As an example of this process, consider the query mention "Tony Blair." Suppose the English document is searched for occurrences of "Tony Blair," which is found to align with the single Arabic word **بلير** for "Blair." By using this projection alone, **بلير** would become the named mention in the target language query. However, sorting all alignments across the entire parallel text by frequency, **توني بلير** turns out to be the most frequent aligned string with "Tony Blair" and that is the Arabic translation of his full name. Since the query document also contains the Arabic "Tony Blair" (which in this case is aligned to the English "Blair,"), the full name will be chosen as the the query name string in this case.

To get a sense for the accuracy of this process, an assessor familiar with the language[11] evaluated the accuracy of automated name projection on a convenience sample of five language pairs. The entire query set was examined for Spanish; for Chinese, Italian, French, and Greek, a random sample of 100 name projections was examined.[12] The evaluators indicated whether the name was completely correct, present in its entirety but including extraneous words, partially correct (meaning that at least one token was correct, but that all required tokens were not present), or completely incorrect. Figure 3 shows the results. The proportion of fully correct projections varies from just under 70% to 98%. Fewer than 10% of the queries were completely wrong for any of the five languages, and no more than 1% of the queries were completely wrong for any Roman script language.

[11] A native speaker, or in the Romance languages, a non-native speaker with years of college study.
[12] See Table 4 for the corresponding ISO-639 two-letter code for each language

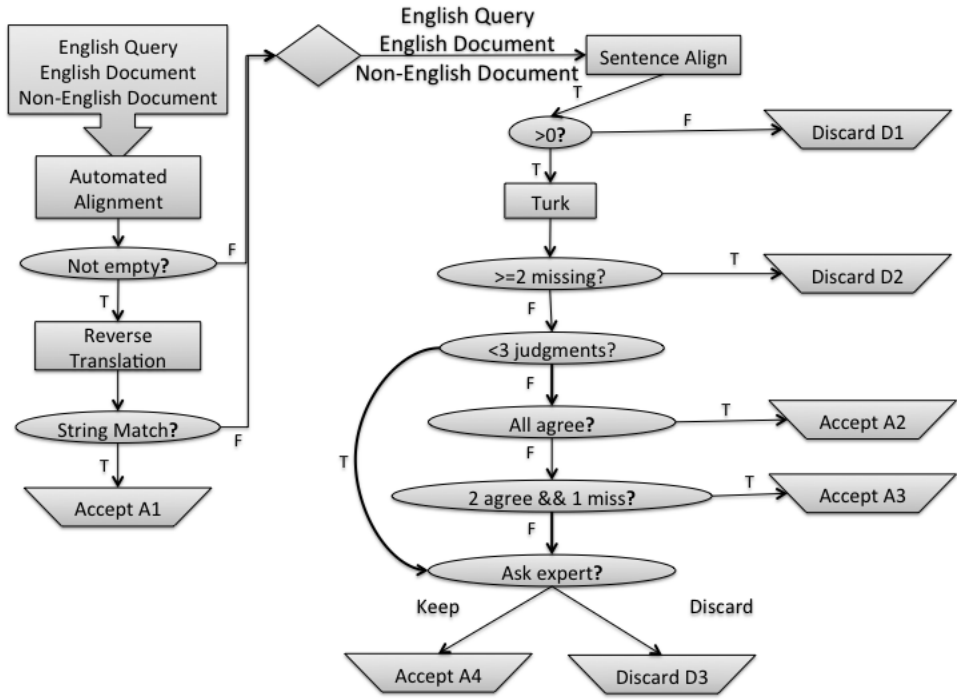

Figure 4: Shows the process followed to determine if a query was accepted. If it was accepted, the category indicates how the name was determined.

Curating Non-English Queries

These results indicate that some way of verifying, and possibly correcting, automated alignments is needed for at least some language pairs. We therefore designed the process shown in Figure 4 for this purpose. The process begins with automated alignment, as described above.

Entirely manual review would be expensive because together our six test collections include more than 50,000 non-English queries. We therefore would prefer to start with some high-precision automated triage process that can automatically accept most reasonable alignments without accepting many unreasonable ones. We do this by using Google Translate[13] to automatically translate each projected name back into English, creating a 'round-trip' translation. If the resulting English translation exactly matches the original English named mention, we treat our projection as correct (Outcome A1 in Figure 4). Like our automatic projection, Google Translate relies on statistical alignments learned from parallel text. As a result, there is some risk that we might accept a bad projection if Google Translate happens to make the same error

[13] `http://translate.google.com/`

(in the other direction). Google Translate is, however, trained on (largely) different parallel texts, and it is optimized for accuracy across all terms, rather than specifically being optimized for coverage of names. We therefore expect incorrect accept decisions based on an exact match of the round-trip translation between two such different systems to be rare, and indeed in the section "Curating Non-English Queries" we show that to be the case. Google Translate currently supports translation from 70 languages to English, including all 21 of the non-English languages used in our test collections. If there is no match (or, in the trivial case, if automated projection yielded an empty query string) then the automatically projected query is sent for manual review and, if necessary, correction.

Crowdsourcing is particularly well suited for tasks that require specialized language skills because a wide variety of language expertise can be found among crowdworkers. In order to create a task for human assessment, the sentence level alignments were presented to Amazon's Mechanical Turkers. There were cases where no aligned sentence was present in our parallel document collection; in such cases, the query was discarded (Outcome D1 in Figure 4). The remaining queries were examined by Turkers.

The Mechanical Turk Human Intelligence Task (HIT) was constructed by first selecting a query, then selecting every English sentence in the query document that had at least one token in common with the query name, and finally presenting that sentence alongside the aligned non-English sentence. Tokens in the English sentence that matched a query token were shown in bold. The Turker was asked to copy and paste the target language characters that best correspond to the English name into a result box. Figure 5 is an example task instance. Because exact string matches in the document were required, Turkers were instructed not to manually enter a "better" name, even if they felt that every name in the non-English sentence shown seemed to them to be deficient in some way. They were also asked to limit their response to a single answer (which could contain multiple tokens). If the name was not present in any non-English sentence, the Turker was instructed to mark "Missing Name." A single HIT consisted of ten tasks like the two shown in Figure 5. Nine of these were for instances where the correct mapping was not known; the tenth was for a known mapping from round-trip translation, which was used to estimate Turker accuracy.

Most of the remaining queries were checked independently by three Turkers. In a few instances, however, responses were obtained from only one or two Turkers. This was in part due to expiring tasks and in part due to an occasional skipped task. If two or more Turkers indicated that the name was missing from the non-English sentences, the query was discarded (Outcome D2 in Figure 4).

For queries with at least three judgments, no more than one of which was "Missing Name" if all Turkers who identified a non-English name agreed on that name, the query was accepted with that name (Outcome A2 in Figure 4 for three agreeing judgments; Outcome A3 for the few cases of two agreeing judges where the third noted the non-English name as missing). The very few queries with fewer than three judgments of any kind, and all queries on which the Turkers exhibited some disagreement about the proper resolution, were

Please identify **Ahmet Sezer** in the Turkish passage.

Ahmet Sezer	PASTE ANSWER HERE ☐Missing Name
President **Ahmet Sezer** has accused the ruling party of trying to penetrate state administration with Islamic ideology.	Cumhurbaşkanı Ahmet Necdet Sezer, iktidar partisini devlet yönetimine İslamcı ideolojiyi sokmaya çalışmakla suçladı.
In turn, Erdogan has criticised **Sezer** for blocking government appointments to public office.	Erdoğan da Sezer'i kamu dairelerine hükümet atamalarının önünü tıkamakla eleştirdi.

Please identify **Goran Kljajevic** in the Turkish passage.

Goran Kljajevic	PASTE ANSWER HERE ☐Missing Name
Among them are former Belgrade Commercial Court president **Goran Kljajevic** and a judge from that court, Delinka Djurdjevic.	Bunlar arasında eski Belgrad Ticaret Mahkemesi başkanı Goran Kljajeviç ve aynı mahkemenin bir hakimi olan Delinka Curceviç de yer alıyor.
Goran Kljajevic's brother, Marko, was the head of the trial chamber in the Zoran Djindjic murder trial.	Goran Kljajeviç'in kardeşi Marko, Zoran Cinciç cinayeti davasındaki hakim kurulunun başkanıydı.
Marko **Kljajevic** withdrew from the trial in late August, objecting to the police and judiciary's treatment of his brother.	Marko Kljajeviç, polis ve yargının kardeşine ettiği muameleye karşı çıkarak Ağustos ayı sonlarında davadan çekildi.

Figure 5: Example Turker Name Projection Tasks

reviewed by someone on our research team who was familiar with at least the character set of the non-English language. This "local language expert" made the final call as to what the correct name projection should be after consulting the full text of the original documents, the automated projection results, and all of the Turker's responses. The local language expert could either select a non-English name string or discard the query.

Partitioning the Test Collection

Finally, we partitioned the resulting queries ({document, name} pairs) into three sets: 60% for training, 20% for development testing (devtest), and 20% for test. Because alignment errors can result in removal of different queries for different non-English languages, even within a multi-way parallel collection, a different number of queries that are available to be partitioned for each language pair. We have built test collections only for language pairs that include English, although by intersecting two collections that are built from the

Table 1: Sources of parallel text

Collection	Obtained from
Arabic	LDC (LDC2004T18)
Chinese	LDC (LDC2005T10)
Europarl5	`http://www.statmt.org/europarl/`
ProjSynd	`http://www.statmt.org/wmt10/`
SETimes	`http://elx.dlsi.ua.es/~fran/SETIMES/`
Urdu	LDC (LDC2006E110)

same multi-way parallel document collection it would be possible to generate a test collection between two non-English languages.

Six Test Collections

For our purposes, we wanted parallel document collections that include a substantial amount of text that is rich in person names, at least some of which refer to well-known people (because the TAC KBP knowledge base that we link to is populated principally with well-known entities). Our interest in relatively large document collections arises from our goal of supporting not just evaluation but also training of entity linking systems that rely on machine learning. Moreover, machine learning systems for cross-language tasks might well themselves require substantial amounts of parallel text for training. We have, therefore, sought parallel document collections that are large enough to make it possible for us to define training, development test, and evaluation partitions. Moreover, multi-way parallel document collections offer the potential for increased leverage by allowing the same ground truth English annotations to be projected to more than one non-English language. We also would like our test collections to include several language families and a diversity of character sets, since such factors can influence the difficulty of the cross-language entity linking task.

As shown in Table 1, we selected six parallel document collections that together satisfy these desiderata and that together pair 21 non-English languages with English. Three of the document collections are multi-way parallel, spanning a total of eighteen non-English languages. To extend the diversity of character sets, we elected to include three bilingual parallel document collections that pair English with Arabic, Chinese, or Urdu.[14] Together, these document collections contain 196,717 non-English documents. As Table 2 shows, these languages span ten language families and five scripts.

[14]The EMEA and EMILLE corpora were also considered, but we concluded that they had inadequate coverage of person names for our purposes.

Table 2: Language characteristics

Language	Code	Source	Script	Family
English	en	all	Latin	Germanic
Arabic	ar	LDC Arabic	Arabic	Semitic
Chinese	zh	LDC Chinese	Traditional	Chinese
Urdu	ur	LDC Urdu	Arabic	Indo-Aryan
Danish	da	Europarl	Latin	Germanic
Dutch	nl	Europarl	Latin	Germanic
Finnish	fi	Europarl	Latin	Uralic
Italian	it	Europarl	Larin	Romance
Portuguese	pt	Europarl	Latin	Romance
Swedish	sv	Europarl	Latin	Germanic
Czech	cs	ProjSynd	Latin	Slavic
French	fr	ProjSynd	Latin	Romance
German	de	ProjSynd	Latin	Germanic
Spanish	es	ProjSynd	Latin	Romance
Albanian	sq	SETimes	Latin	Albanian
Bulgarian	bg	SETimes	Cyrillic	Slavic
Croatian	hr	SETimes	Latin	Slavic
Greek	el	SETimes	Greek	Greek
Macedonian	mk	SETimes	Cyrillic	Slavic
Romanian	ro	SETimes	Latin	Romance
Serbian	sr	SETimes	Latin	Slavic
Turkish	tr	SETimes	Latin	Turkic

Process Statistics

In this section, we draw on our experience with these six test collections to characterize the process described above in section on building test collections.

Generating English Queries

Across the six parallel document collections, English Named Entity Recognition (NER) identified 257,884 named references to people that were unique within an English document. Applying our four filtering rules (single-token deletion, single-instance deletion, substring containment deletion, and no more than 10 instances for any query string) resulted in 20,436 English queries for the Turkers to check.

Generating Ground Truth Links

We collected three judgments for each of the 20,436 queries. A total of 314 unique Turkers contributed to at least one batch, with 165 unique Turkers contributing to more than one batch. As Figure 6 shows, the

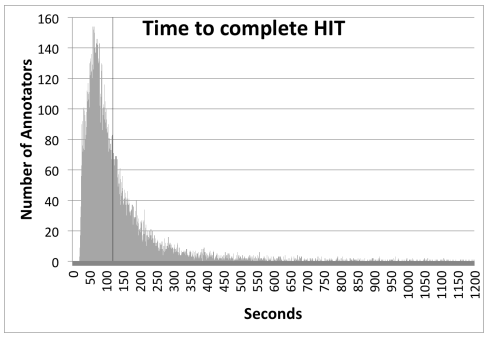

Figure 6: Distribution of the time Turkers spent on task, in seconds. The vertical bar at 120 seconds visually indicates that the task typically required less than two minutes.

average time required to complete a HIT (i.e., six queries, two of which were known items) was 2.5 minutes, and the mode was well below two minutes (marked by a black line for reference). The long tail is most likely due to Turkers who were distracted in the middle of completing a task (we allowed Turkers to take as much as one hour to finish a HIT). Only 14 of the 314 Turkers averaged under one minute per HIT, and no assessments were eliminated due to insufficient time spent on the task.

Turker disqualification for accuracy below 80% on known items resulted in the removal of 3,095 queries (22%) across the six test collections. Turker disagreement resulted in a similar deletion rate, 3,018 queries (17%). Only in 362 cases (2%) did Turkers indicated that NER had produced a query that was not actually a named reference to a person, and in only 14 additional cases (<1%) did any Turker indicate that they were unable to decide. In 1,010 cases (7%), more than one of these reasons resulted in deletion of a query. These deletions resulted in retention of a total of 14,957 Turker-resolved queries across the English parts of the six parallel document collections.

As Figure 7 illustrates, Turkers who achieved lower accuracy on the known items in a batch were more likely to have their queries removed from that batch for other reasons. In that figure, the difference between the black and gray vertical bar indicates the number of removed queries for Turkers with the indicated (binned) level of computed accuracy. Below 80%, all queries were removed, above that value progressively fewer queries needed to be removed (for any reason) with increasing Turker accuracy. Also notable is that the number of judgments provided by a Turker was strongly correlated with computed accuracy (note the log scale). From this we conclude that our 80% threshold seems to have been an appropriate choice, likely resulting in relatively few unnecessarily removed queries.

The mean Turker accuracy over all batches was 94.6% with a standard deviation of 11.3%; for 647 of the 1,139 Turker-batches, the Turker achieved a perfect score. A total of 81 of Turker-batches were eliminated for accuracy below 80%; however, the number of HITs submitted by these Turkers was very low, apparently

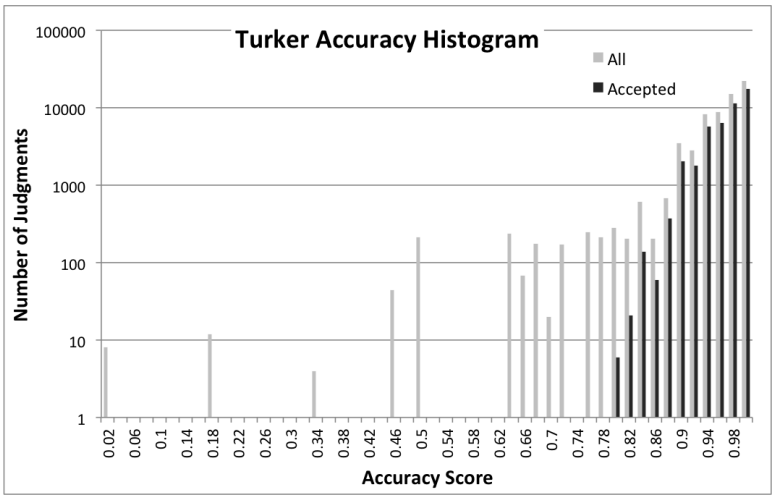

Figure 7: Histogram of Turker accuracy by judgment (6 per HIT).

because most poor performers only sampled the task. From this we conclude that our accuracy checks were adequate to discourage spammers and bots (who were not paid if they did not achieve 80% accuracy on known items).

The resulting English queries exhibit relatively little ambiguity, with only 13 of the 3,433 unique English query strings that occur between 2 and 10 times having distinct queries with the same query string that resolve to two or more different entities. That is not to say that these strings are unambiguous in Wikipedia; indeed, many are not. But it does indicate that there is a strong single-sense-per-test-collection phenomenon at work for person name entities in these test collections, as might be expected.

Generating Non-English Queries

The number of English queries for a particular parallel test collection sets the upper bound for the number of non-English queries for any other language in that test collection, and projection failures (in which the non-English string is empty) further limit the number of non-English queries that can be considered for inclusion in any given non-English test collection. As Figure 8 shows, on average about 10% of English queries yield projection failures (ranging from 4% in Albanian to 30% in French). When viewed by document collection, Project Syndicate was the only one to systematically exhibit an unusually large projection failure rate (ranging from 15% for German to 30% for French). Much of that failure rate is attributable to incomplete parallelism in the multi-language document collection. In Project Syndicate, we used the English side of parallel Czech-English document pairs as a basis for query formulation in English, but it turned out that several of those English documents had no corresponding French document; in some other cases a less

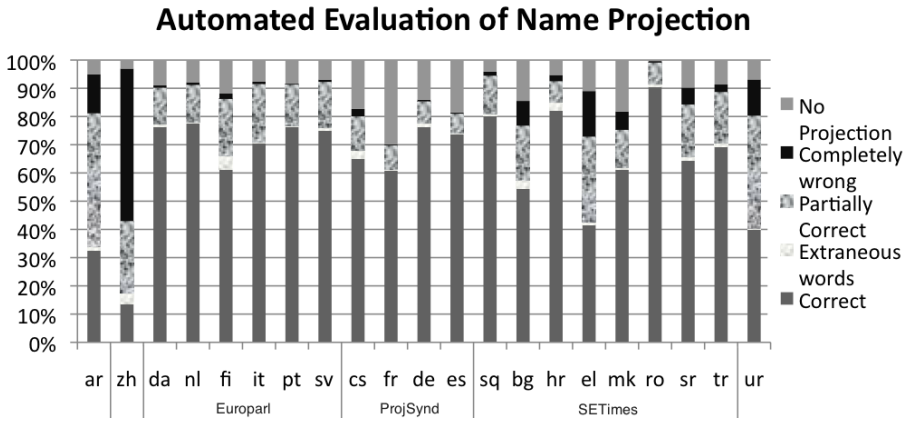

Figure 8: Automated evaluation of name projection when Google Translate is used to translate the projected queries back into English.

complete French document existed, and that French document lacked the portion that contained the query mention. As a result, successful projection was possible for more Czech than for French queries. However, since we start with over 1,000 queries on the English side of every parallel document collection, loss rates of this magnitude are easily tolerated.

Round-trip translation of the non-empty queries yielded an exact match with the original multi-token English query in 72% of the cases.[15] The exact match rate exhibits quite substantial variation by language, however, with Chinese exhibiting the lowest exact match rate (14%) and Romanian the highest (91%). Overall, about a quarter of the non-English queries required manual curation. Of queries requiring manual curation, in 93% of the cases the round-trip translation and the original English query share at least one token in common. From this we conclude that our approach to manual curation is well focused on reasonable candidates, and thus reasonably likely to achieve a useful yield.

Curating Non-English Queries

We personally performed complete expert review for Chinese and Spanish. For the remaining 19 non-English languages, 98 different Turkers participated in curation of the non-English queries, completing an average of 46 HITs (minimum 1, maximum 302). The fastest HIT was completed in 42 seconds, but the median amount of time spent on the task was just shy of two minutes, while the average was about two and a half minutes. About two-thirds of the Turkers achieved above 90% accuracy on queries with known ground truth (which in every case were queries that had yielded exact match round-trip translations).

[15]The percentages here are of non-empty projected queries; multiply by 0.9 to obtain the fraction of the original English queries shown in Figure 8.

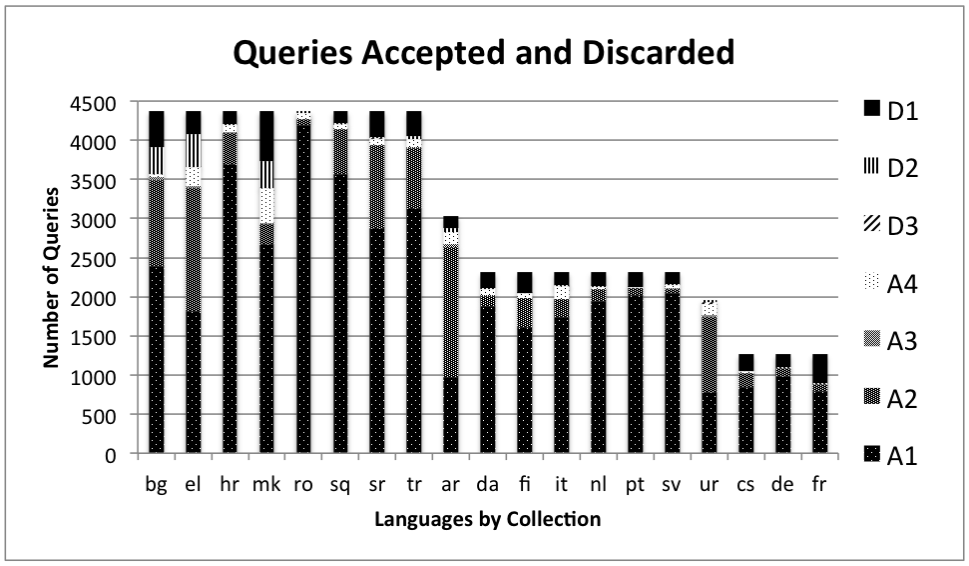

Figure 9: Following the flow-chart in Figure 4, this chart shows the number of queries that were accepted or discarded for each reason

Figure 9 summarizes number of queries that were accepted or discarded in each of the stages depicted in Figure 4. Illustrating the best results, very few Romanian queries were discarded for any reason. This is in part due to the fact the multi-way parallel document collections are not equally complete in every language. In this case, the English documents from which we built our queries for the South-East European Times document collection all had Romanian counterparts, whereas many of these same documents lacked, for example, Macedonian counterparts.

One notable pattern is that round-trip translation (accept point A1 in Figure 4) yielded a much higher proportion of the total accepted queries for Roman script languages (79%) than for non-Roman script languages (48%, averaged over Arabic, Bulgarian, Chinese, Greek, Macedonian, and Urdu). Manual curation of queries that failed round-trip translation generally leveled the playing field, however, to 87% and 85%, respectively. This indicates the manual curation is much more important for non-Roman script languages.

Because we accept exact matches after round-trip translation as correct without any manual review, it is important that we characterize the false positive rate of this test. One way of doing this is to examine cases in which we have three Turkers who "corrected" known items that happened to be exact match round-trip translations. If the majority response was the same as the round-trip translation, we counted round-trip translation as correct. We performed this analysis for the 19 non-English languages for which Turkers provided judgments. On average, this yielded 92% accuracy for round-trip translation (and over 95% for

Table 3: Fraction of all person names lost as queries due to various factors during the query creation phase

Reason for Attrition	Queries Lost
Single-word name	45.1%
Only one occurrence of name in collection	15.8%
Ten occurrences of name already included	11.6%
Manual name curation	5.0%
To avoid predicted NIL/non-NIL imbalance	4.0%
More descriptive name appears in document	1.1%
Low Turker accuracy	0.9%
Turker disagreement	0.9%
Could not locate name in English document	0.5%
Missing judgments	0.3%

13 of the 19 languages). Indeed, four languages (Italian, French, Swedish and Romanian) achieved 100% round-trip translation accuracy by this measure. The two languages for which round-trip translation was least accurate were Arabic (64%) and Finnish (55%), but for Finnish inspection of the discrepancies indicated that most were due to the fact that Turkers excluded morphemes at the end of the names (Finish exhibits morphological variation on names) while the round-trip translation preserved those morphemes. If we were to (manually) consider morphological variants as equivalent, the accuracy of round-trip translation on Finish would have been measured at 91%. From this analysis, we contend that the error rate for our Arabic test collection may be unacceptably high, but that for the other languages we would expect the overall error rate from this source to be perhaps 5%. It seems reasonable to expect that error rates that low would have little effect on relative comparisons across systems within a language, or across languages for systems of similar design.

Test Collection Statistics

The various sources of query attrition, together with the percentage of the person names lost for each, are shown in Table 3. With additional work, some of these forms of attrition might be ameliorated, but none seem unreasonably large. A total of 14,806 English queries resulted from this process. These correspond to 59,224 queries across the 21 non-English languages. Further attrition caused by projecting the English names into those 21 languages resulted in an initial non-English query count of 55,244. Some of the error due to misalignment was rectified during the hand-curation process, although strict requirements for agreement lead to an overall drop in the number of hand-curated queries. That query set contains a total of 54,819 queries. Ultimately, the full set of non-NIL queries resolved to a total of 1,618 distinct Wikipedia entities.

Because detecting when an entity cannot be resolved is an important requirement in many entity linking applications, we would like to have a substantial number of NIL queries (i.e., those for which no resolution can be made), but we do not want so many NIL queries that just guessing NIL all the time would be a viable strategy. As Table 4 shows, the fraction of the queries that resolve to NIL varies between 26% and 51% for every language except Arabic (for which 88% of queries resolve to NIL). Of course, users of the test collection could sample the queries in a way that is stratified on the NIL/non-NIL distinction to achieve any desired ratio.

Table 4: Language coverage in the test collections

	Document	**Machine Translation**		**Hand Curation**	
Language	**Collection**	**Queries**	**Non-NIL**	**Queries**	**Non-NIL**
Arabic (ar)	LDC Arabic	2,916	679	2,867	642
Chinese (zh)	LDC Chinese	1,958	956	1,996	970
Urdu (ur)	LDC Urdu	1,828	1,093	1,907	1,140
Danish (da)	Europarl	2,105	1,096	2,093	1,092
Dutch (nl)	Europarl	2,131	1,087	2,128	1,086
Finnish (fi)	Europarl	2,038	1,049	2,042	1,051
Italian (it)	Europarl	2,135	1,087	2,139	1,089
Portuguese (pt)	Europarl	2,119	1,096	2,122	1,096
Swedish (sv)	Europarl	2,153	1,107	2,154	1,106
Czech (cs)	ProjSynd	1,044	722	1,045	723
French (fr)	ProjSynd	885	657	889	658
German (de)	ProjSynd	1,086	769	1,091	774
Spanish (es)	ProjSynd	1,028	743	1,028	743
Albanian (sq)	SETimes	4,190	2,274	4,204	2,285
Bulgarian (bg)	SETimes	3,737	2,068	3,572	1,965
Croatian (hr)	SETimes	4,139	2,257	4,186	2,280
Greek (el)	SETimes	3,890	2,129	3,659	1,999
Macedonian (mk)	SETimes	3,573	1,956	3,382	1,837
Romanian (ro)	SETimes	4,355	2,368	4,345	2,361
Serbian (sr)	SETimes	3,943	2,156	4,013	2,195
Turkish (tr)	SETimes	3,991	2,169	4,018	2,183
Total		55,244	29,518	54,819	29,275

Amortizing what we paid to all of our Turkers over the full set of non-English queries, the cost works out to about 10 cents per query. Notably, however, that estimate does not include our own effort to manage the process, nor the cost of developing the parallel document collections which we used.

Threats to the Test Collection's validity

There are a few threats to the validity of these test collections, four of which will be discussed herein. The first has to do with the manner in which query names were determined. This was done using named entity recognition followed by a light curation where an author looked over the names and removed ones that did not appear to represent person names. It is possible that companies which incorporate a person's name could have been included, while references to people such as "Iron Lady" may have been excluded. Instances of the former were likely excluded during a subsequent phase, while instances of the latter could make the entity linking task easier given that a system would not have to cope with such names.

A second threat is introduced by using an English entity linking system to propose candidates for linking. Although we expect the miss-rate where the entity appears in the knowledge base but is considered to by be NIL in the collection to be small, it also means the we expect high performance on the monolingual task. Thus the challenge presented by the test collections comes almost entirely from the cross lingual aspect of the problem.

The use of parallel collections in the solution to the entity linking problem represents a third threat to the validity of the test collection. It was assumed by the authors that solutions would not make use of the parallel English text as it is a cornerstone in the creation of the collection.

Finally, the focus on Wikipedia for resolvable entities represents a fourth threat to validity. In particular it means that some potential applications of the test collection are better than others. This creates a threat to generalizability.

Building Cross-Language Entity Linking Systems

With these test collections, we can now begin to characterize the extent to which cross-language entity linking techniques can be easily adapted to new languages. We do this by first extending our existing monolingual entity linking system to accommodate the additional challenges of cross-language entity linking and then automatically adapting that extended system to the specifics of each language pair. The approach used for *monolingual* entity linking breaks the problem into two phases: (1) identification of a relatively small set of plausible KB entities, or *candidate identification*; and (2) ranking of those candidates using supervised machine learning (*candidate ranking*). The ranking step orders the candidates, including NIL, by the likelihood that each is a correct match for the query entity. Table 5 illustrates a few representative Turkish queries. Clearly, both candidate identification and candidate ranking will be somewhat more complex in cross-language entity linking than in the same-language case.

For candidate generation, our same-language entity linking system uses a number of quickly calculable name comparisons. Indexes are created to support rapid identification of (1) KB entries with an exact

Table 5: Example queries

Turkish Query	Document Excerpt	KBID/NIL	KB Title
Hoe Biden	Karar, ABD Başkan Yardımcısı **Hoe Biden**'ın BH'ye yapacağıziyaret öncesinde çıktı.	E0747316	Joe Biden
Rajko Daniloviç	Ancak Cinciç ailesinin avukatı **Rajko Daniloviç**, Lukoviç'i kimin koruduğunun bilinmesinin önemli olduğunu söyleyerek buna karşı çıkıyor.	NIL	
Haris Silaciç	Ancak dört yıl önce yapılan Boşnak cumhurbaşkanlı ğı üyesi yarışını az farkla ikinci sırada tamamlayan **Haris Silaciç**, değişikliklere karşsı çıkıyor ve büyük bir destekçi kitlesine sahip bulunuyor.	E0305255	Haris Silajdžić

name match; (2) entities with an alternative name that matches the query (e.g., *Duchess of Cambridge* for *Catherine Middleton*); (3) entities with name fragments (given names or surnames) in common with the query; and (4) entities sharing character 4-grams with the query entity. This candidate identification phase provides three to four orders of magnitude of reduction in the number of entities to which the full battery of comparison features must be applied. Candidates are then ranked using a ranking support vector machine (SVM^{rank}) [20]. Feature vectors representing candidate alignments to knowledge base entries include features based on name similarity, textual context, matches of relations found in the knowledge base, named entities that occur in both the knowledge base and the query document, and indications of absence from the knowledge base. A detailed description of the entity linking system can be found in McNamee et al.[27].

To construct cross-language entity linking systems, we augmented our monolingual entity linking system with features based on transliteration (for name matching) and cross-language information retrieval (for matching terms surrounding the name mention against terms found in a candidate entity's English Wikipedia page). We learned both types of features for each language by leveraging the parallel text in the training partition as described in [29]. Feature combination was tuned for each language on the devtest partition, and the results in Table 6 for each language were evaluated on the test partition. Our general approach proved not to work well for Chinese because our statistical transliteration model that was learned from parallel text was not designed for the large fanout in the many-to-one relations necessary to align English and Chinese characters. We therefore report results only for the other 20 non-English languages.

Table 6 reports micro-averaged accuracy of the top-ranked result for each language, including both NIL and non-NIL entities (in the ratio that they appear in the test collection). The English column shows the same measure for the corresponding English queries as a monolingual baseline. Table 6 also illustrates the effect of manual curation for name projection on the results. Comparing the manually curated results with the results from only the fully automatic alignment earlier in our process (i.e., the input to manual

Table 6: Success@1 on test queries – English Queries and analogous cross-language queries

Language	English	Hand-Curated Cross-Language		Machine-Aligned Cross-Language	
Arabic	0.9480	0.9263	(97.7%)	0.8908	(94.0%)
Bulgarian	0.9818	0.9350	(95.2%)	0.8922	(90.9%)
Czech	0.9310	0.8585	(92.2%)	0.8325	(89.4%)
Danish	0.9883	0.9789	(99.0%)	0.9556	(96.7%)
German	0.9309	0.9128	(98.1%)	0.8479	(91.1%)
Greek	0.9794	0.8840	(90.3%)	0.8492	(86.7%)
Spanish	0.9087	0.8750	(96.3%)	0.8846	(97.3%)
Finnish	0.9859	0.9368	(95.0%)	0.9412	(95.5%)
French	0.9301	0.8930	(96.0%)	0.8548	(91.9%)
Croatian	0.9799	0.9497	(96.9%)	0.9374	(95.7%)
Italian	0.9842	0.9009	(91.5%)	0.9187	(93.3%)
Macedonian	0.9778	0.8950	(91.5%)	0.8542	(87.4%)
Dutch	0.9841	0.9751	(99.1%)	0.9478	(96.3%)
Portuguese	0.9865	0.9269	(94.0%)	0.9774	(99.1%)
Romanian	0.9761	0.9738	(99.8%)	0.9658	(98.9%)
Albanian	0.9717	0.9191	(94.6%)	0.9105	(93.7%)
Serbian	0.9762	0.8370	(85.7%)	0.8335	(85.4%)
Swedish	0.9866	0.9710	(98.4%)	0.9665	(98.0%)
Turkish	0.9801	0.9642	(98.4%)	0.9677	(98.7%)
Urdu	0.9725	0.8763	(90.1%)	0.8264	(85.0%)
Average	0.9680	0.9195	(95.0%)	0.9027	(93.2%)

curation, with all of its mistakes) indicates that manual curation actually yields better results for 15 of the 20 languages. We would expect that applying the manual curation process to the test partitions would reduce the variance of our evaluation measure somewhat, but the observed systematic improvements in accuracy that we observe seem to us to be more likely to result principally from from manual curation of the queries in the training and devtest partitions than to result from more accurate measurement alone.

On average across the 20 languages, the cross-language system achieves 95% of the monolingual English baseline by this measure. Indeed, only for one language (Serbian) does the cross-language system achieve less than 90% of the monolingual baseline. Looking more closely for patterns, we see that cross-language entity linking accuracy is nearly as good as the monolingual English baseline for Arabic, German and Spanish, but Bulgarian and Greek do relatively poorly. The Bulgarian and Greek results comport with a hypothesis that language pairs that require transliteration pose additional challenges for cross-language entity linking, and we suspect that the large number of NILs in Arabic may have masked any transliteration effect that

would have been present in that language. Such a hypothesis is not sufficient to explain the relatively poor performance on Serbian, however.[16]

Perhaps the most important pattern, however, is that the absolute results are quite good. Always guessing the most common result (NIL) would yield a Success at Rank 1 (S@1) value below 0.51 for every language except Arabic, and even our lowest absolute result is substantially above that. Having shown that we can link entities from news stories in 20 languages to Wikipedia with better than 86% top-1 accuracy seems well worth the effort, and this is something that we could not as easily have shown without the method for affordably creating useful test collections that we have described in this paper.

Conclusion

Our principal goals in this paper have been to describe an affordable way of creating cross-language entity linking test collections and to apply that method to create several such test collections for a wide range of languages. We hope in this way to help researchers to identify which aspects of their approaches to cross-language entity linking work well in many language pairs and which are language specific, to promote the development of language-neutral approaches to cross-language entity linking that will be applicable to many of the world's languages, and to foster entity linking research by researchers who have interest in a specific language that is not currently supported by existing test collections. Beyond merely demonstrating techniques, we are also sharing the six test collections that we have built.

Looking to the future, there is of course more to be done. Perhaps most obviously, the recently-released Europarl v7 should permit development of test collections for additional Central and Eastern European languages. The existence of several languages in a single test collection might also be exploited in novel ways, for example to evaluate mixed-language entity linking. Of course, our comparative analyses in this paper only scratch the surface of what might be done along those lines. And our challenges with Arabic and Chinese clearly indicate that there is still room for careful development of language-specific techniques in particular cases.

Thinking even more broadly, we have illustrated that Mechanical Turk is more than a source of labor; in our case it was an essential source of skilled labor, labor with a diverse set of language skills that we could not have as easily obtained in any other way. Combining statistical techniques with crowdsourced human intelligence in ways that make use of to the strengths of both can help both to control costs and enhance quality, and we expect such combinations to become increasingly common in the coming years.

[16] Serbian can be written in either the Latin or Cyrillic alphabet, but in the South-East European Times document collection that we used Serbian is written in the Latin alphabet.

Acknowledgments

We are grateful to the many Mechanical Turk annotators who provided us with fast, accurate responses to our requests. Curation assistance was freely provided by Tan Xu, Mohammad Raunak, Mossaab Bagdouri, Árpád Beszédes, Veselin Stoyanov, Ravandar Lal, and Damianos Karakos, for which we are grateful. We are also grateful to the creators of the Europarl [23] and South-East European Times document collections [43], and to Chris Callison-Burch and Ann Irvine for their support with machine translation and orthographic transliteration.

References

[1] Sisay Fissaha Adafre and Maarten de Rijke. Discovering missing links in Wikipedia. In *LinkKDD '05: Proceedings of the Third International Workshop on Link Discovery*, pages 90–97, New York, NY, USA, 2005. ACM.

[2] Amazon.com. Amazon Mechanical Turk, 2005.

[3] Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigo. Overview of theWeb People search clustering and attribute extraction tasks. In *CLEF Third WEPS Evaluation Workshop*, 2010.

[4] Krisztian Balog, Leif Azzopardi, Jaap Kamps, and Maarten De Rijke. Overview of WebCLEF 2006. In *Cross-Language Evaluation Forum*, pages 803–819, 2006.

[5] Alex Baron and Marjorie Freedman. Who is Who and What is What: Experiments in cross-document co-reference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 274–283, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

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

[7] Chris Callison-Burch and Mark Dredze. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[8] Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra. Exploiting context analysis for combining multiple entity resolution systems. In *Proceedings of the 35th SIGMOD international conference on Management of data*, SIGMOD '09, pages 207–218, New York, NY, USA, 2009. ACM.

[9] Cyril W. Cleverdon. The Aslib Cranfield research project on the comparative efficiency of indexing systems. *ASLIB Proceedings*, 12(12):421–431, 1960.

[10] Cyril W. Cleverdon. The Cranfield tests on index language devices. *ASLIB Proceedings*, 19(6):173–194, 1967.

[11] Andrew W. Cole. Corpus development and publication. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1825–1830, May 2006.

[12] Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, and Jonathan Wright. Linguistic resources for 2012 knowledge base population evaluations. In *Text Analysis Conference (TAC)*, 2012.

[13] Meghan Lammie Glenn, Stephanie Strassel, Lauren Friedman, Haejoong Lee, and Shawn Medero. Management of large annotation projects involving multiple human judges: A case study of GALE machine translation post-editing. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[14] Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 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 of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 923–931, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[15] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[16] Lynette Hirschman. The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech & Language*, 12(4):281–305, 1998.

[17] Ann Irvine, Chris Callison-Burch, and Alexandre Klumentiev. Transliterating from all languages. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, 2010.

[18] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, June 2011. Association for Computational Linguistics.

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

[20] Thorsten Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA, 2006. ACM.

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

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

[23] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005.

[24] Dawn Lawrie, James Mayfield, Paul McNamee, and Douglas W. Oard. Creating and curating a cross-language person-entity linking collection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 2012.

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

[26] Inderjeet Mani, Alex Yeh, and Sherri Condon. Learning to match names across languages. In *Multi-source Multilingua Information Extraction and Summarization Workshop (MMIES)*, pages 2–9. Association for Computational Linguistics, 2008.

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

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

[29] Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W. Oard, and David Doermann. Cross-language entity linking. In *International Joint Conference on Natural Language Processing (IJCNLP-2011)*, Chiang Mai, Thailand, November 2011. Association for Computational Linguistics.

[30] Julie Medero, Kazuaki Maeda, Stephanie Strassel, and Christopher Walker. An efficient approach to gold-standard annotation: Decision points for complex tasks. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2460–2463, May 2006.

[31] Christian Meilicke, Raul Garcia-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Andrei Tamilin, Cássia Trojahn dos Santos, and Shenghui Wang. MultiFarm: A benchmark for multilingual ontology matching. *Journal of Web Semantics*, 15:62–68, 2012.

[32] Keith J. Miller, Elizabeth Schroeder Richerson, Sarah McLeod, James Finley, and Aaron Schein. International multicultural name matching competition: Design, execution, results, and lessons learned. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3111–3117, 2012.

[33] David N. Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518, 2008.

[34] Vincent Ng. 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, 2010.

[35] Jian-Yun Nie. *Cross-Language Information Retrieval*. Synthesis Lectures on Human-Language Technologies. Morgan and Claypool, 2010.

[36] Lev Ratinov and Dan Roth. 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 2009. Association for Computational Linguistics.

[37] Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative research: Distributed Computing - Volume 2*, CASCON '93, pages 1071–1082, 1993.

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

[39] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[40] Ling-Xiang Tang, In-Su Kang, Fuminori Kimura, Yi-Hsun Lee, Andrew Trotman, Shlomo Geva, and Yue Xu. Overview of the NTCIR-10 cross-lingual link discovery task. In *Proceedings of the 10th NTCIR Conference*, pages 1–31, 2013.

[41] Barbara Tillett. Authority control: State of the art and new perspectives. In *International Conference Authority Control: Definition and International Experiences*, 2003. http://eprints.rclis.org/4193/.

[42] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Conference on Natural Language Learning (CoNLL)*, 2003.

[43] Francis Tyers and Murat Serdar Alperen. South-East European Times: A parallel corpus of Balkan languages. In Stelios Piperidis, Milena Slavcheva, and Cristina Vertan, editors, *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valetta, Malta, 2010.

[44] Jianqiang Wang and Douglas W. Oard. Matching meaning for cross-language information retrieval. *Information Processing and Management*, 48(4):631–653, 2012.

[45] David Yarowsky and Grace Ngai. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Annual Conference of the North Americal Chapter of the Association for Computational Linguistics (NAACL)*, 2001.