Creating and Curating a Cross-Language Person-Entity Linking Collection

Dawn Lawrie†‡, James Mayfield†, Paul McNamee†, Douglas W. Oard†§

†Johns Hopkins University Human Language Technology Center of Excellence
‡Loyola University Maryland
§College of Information Studies and UMIACS, University of Maryland, College Park

Abstract

To stimulate research in cross-language entity linking, we present a new test collection for evaluating the accuracy of cross-language entity linking in twenty-one languages. This paper describes an efficient way to create and curate such a collection, judiciously exploiting existing language resources. Queries are created by semi-automatically identifying person names on the English side of a parallel corpus, using judgments obtained through crowdsourcing to identify the entity corresponding to the name, and projecting the English name onto the non-English document using word alignments. Name projections are then curated, again through crowdsourcing. This technique resulted in the first publicly available multilingual cross-language entity linking collection. The collection includes approximately 55,000 queries, comprising between 875 and 4,329 queries for each of twenty-one non-English languages.

Keywords: Entity linking, Multi-lingual, Crowdsourcing collection building

1. Introduction

Given a mention of an entity in a document and a set of known entities, the entity linking task is to find the entity ID of the mentioned entity, or return NIL if the mentioned entity was previously unknown. In the cross-language entity linking task, the document in which the entity is mentioned is in one language (e.g., Turkish) while the set of known entities is described using another language (in our experiments, English). Entity linking is a crucial requirement for automated knowledge base population.

Entity linking has been the subject of significant study over the past five years. Pioneering work focused on matching entity mentions to Wikipedia articles (Bunescu and Pasca, 2006; Cucerzan, 2007). Although focused on clustering equivalent names rather entity linking, the ACE 2008 workshop conducted evaluations of cross-document entity coreference in Arabic and English (Baron and Freedman, 2008) but not across languages. In 2009, the Text Analysis Conference (TAC) Knowledge Base Population track (TAC KBP) conducted a formal evaluation of English entity linking using a fixed set of documents and Wikipedia articles (McNamee and Dang, 2009). Shared tasks with a variety of characteristics have since emerged elsewhere, including CLEF (Artiles et al., 2010), FIRE (Tiwari et al., 2010), and NTCIR. Very recently, TAC and NTCIR have both for the first time defined a shared task for cross-language entity linking.

The goals of this work are to identify a way to efficiently create and curate cross-language entity linking training and test data and to apply that method to create such collections in many languages. We hope by doing this to accelerate the identification of the best methods for performing cross-language entity linking; to foster entity linking research by researchers who have interest in specific languages beyond the few languages that are supported by existing evaluations; and to promote the development of language-neutral approaches to cross-language entity linking that will be applicable to many of the world’s languages.

This work produces a set of queries in each target language. A query consists of a query id, a string representing the entity, a document ID indicating the document that contains the entity, the type of entity, and the knowledge base entity id (or NIL for entities not found in the knowledge base). The knowledge base is the TAC knowledge base, which is derived from an October 2008 subset of Wikipedia pages that contained Infoboxes; it includes more than 114k persons. This format matches the format of the TAC query sets. Example Turkish queries appear in Table 1.

This paper reviews the methodology we used to create the test collection in Section 2. It then gives a detailed account of the curation of bilingual name alignment in Section 3. Section 4. reports interesting statistics from the resulting collection.

2. Collection Creation Overview

Our approach to collection creation has two distinguishing characteristics: the use of parallel document collections to allow most of the work to occur in a single language; and the use of crowdsourcing to quickly and economically generate many human judgments. A fundamental insight on which the 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.

As an overview of the process, we apply English named entity recognition (NER) to find person names in text, an English entity linking system to identify candidate entity IDs, and English annotators to select the correct entity ID for each name. Standard statistical word alignment techniques are used to map from name mentions in English documents to the corresponding names in non-English documents. Finally, crowd-sourcing is used again to curate the name projections. The increasing availability of multi-way parallel
We used the HLTCOE entity linking system (McNamee et al., 2009) to create a ranked list of candidate entities from the TAC KBP knowledge base and presented the top three entries to human judges. We collected human judgments using Amazon’s Mechanical Turk (2005), which has been applied to a wide array of HLT problems (Snow et al., 2008; Callison-Burch and Dredze, 2010). A paid assessor, called a ‘Turker,’ could select one of the three candidates, “None of the above” (if none of the three was the correct referent), “Not a person” (indicating an NER error) or “Not of the above” (if none of the three was the correct referent).

As an example of this process, consider the query “Joe Biden.” The English document is searched for occurrences of “Joe Biden,” and through word alignment, it is found to align in the Turkish document with “Biden.” By using the projection alone, the Turkish query would become “Biden;” however, by using the collection information, the most frequent alignment of “Joe Biden” is “Hoe Biden” in Turkish. The query document also contains “Hoe Biden,” which aligns to the English ‘Biden.’ Because the projection process chooses the most frequent alignment in the collection, “Hoe Biden” is selected as the string to represent the query. To estimate the accuracy of this approach to name projection, we translated all the Turkish names back into English using Google Translate\(^3\) and compared the results with the

3http://translate.google.com/

and only if all three Turkers agreed on the answer. More details about this process appear in Mayfield et al. (2011).
original English query set. If we found an exact string match between the two English names, we considered the Turkish name to be correct.4 Of the 4,370 English queries, 379 had no projection in the Turkish parallel text. When judging the accuracy of the remaining 3,991, 76% had an exact match with the Google Translated name. Of those remaining, 794 partially matched, 47 had extraneous words, and 116 were completely different.

One could limit the collection to the three thousand queries that Google Translate identified as correct; however, nearly 78% of those are cases where the English and Turkish names appear exactly the same. Reliance on only these queries would create a bias towards entity linking systems that are based simply on name matching, and hinder research that would address more complicated queries. On the other hand if one used all the machine aligned queries, a few bad queries could reduce the usefulness of the collection. We therefore asked Amazon’s Mechanical Turkers to evaluate 1,336 such queries, which are those that failed to have an exact Google translate match and those where name projection failed. The Turkers were asked to examine machine-aligned sentences. Sentences from the document were selected if they included any part of the name of interest. The English name was highlighted in bold as shown in Figure 1. The instructions asked the Turker to copy and paste the Turkish characters that best correspond to the English name. If the name was not present in the Turkish text, the Turker was instructed to mark “Missing Name.” Because exact string matches in the document were required, they were also told not to manually enter a better name that did not appear in the Turkish text. Finally, they were asked to paste only one version of the name.

A work unit consisted of ten tasks like the two shown in Figure 1. Nine of these were for instances where the name was unknown and the tenth was for a known name-projection failure. The Turkers were asked to examine machine-aligned sentences. Sentences from the document were selected if they included any part of the name of interest. For this resolution, the assessor was asked to choose among the options provided by the Turkers. There were usually two choices identified by the Turkers, with either the longest name or highest vote-getter being the correct option. An assessor familiar with the writing system could make accurate decisions given the similarity of person names across languages.

Eight different Turkers participated in the Turkish task. The number of work units undertaken by a particular Turker ranged from two to 103; the average number of tasks was 40.5. Most Turkers scored above 95% on queries with known ground truth. The lowest accuracy was 85% over thirteen work units. The fastest work unit was completed in 52 seconds, but on average it took Turkers two and a half minutes. There were four instances where Turkers submitted answers that did not have an exact string match in the document. In three of these cases the Turker eliminated a middle name or distinguishing characteristic not in the English name as in “Başpatriği 1’inci Bartolomeus,” and in the final case an accented character was not copied correctly. Of the 957 queries where Google Translate identified a problem, 83 were changed as a result of this curation step. This underestimate of alignment accuracy based on Google Translate is due in part to the presence of accented characters, which when present in the translated version prevented an exact string match. In addition, 49 of the 379 queries where the Berkeley Aligner failed to find an alignment were restored to the collection through curation.

### 4. Collection Statistics

A desirable characteristic of an entity linking test collection is balance between the number of NIL queries (i.e., those for which no resolution can be made) and non-NIL queries; detecting that an entity cannot be resolved is an important requirement in many entity linking applications. Table 3 shows that this goal was well met.

The NER system originally identified 257,884 English person names across the six parallel collections. Not all of these names end up as queries; significant attrition occurs in an effort to maintain collection quality. The various sources of query attrition, together with the percentage of the person names lost for each, are shown in Table 4. Some of these forms of attrition could be ameliorated to increase the collection size. A total of 14,806 English queries resulted from our procedure. These correspond to 59,224 queries.

---

4This approach does not guarantee query correctness (Google Translate might itself correct errors in the input). In Turkish Google Translate was verified to be 100% accurate in these cases. The language most susceptible to the problem is Chinese, where all names were curated by humans.

5Serbian can be written in both Latin and Cyrillic alphabets; our collection uses the Latin alphabet.

---

Table 3: Language coverage in our collection.

<table>
<thead>
<tr>
<th>Language</th>
<th>Collection</th>
<th>Queries</th>
<th>Non-NIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (ar)</td>
<td>Arabic</td>
<td>2,829</td>
<td>661</td>
</tr>
<tr>
<td>Chinese (zh)</td>
<td>Chinese</td>
<td>1,958</td>
<td>956</td>
</tr>
<tr>
<td>Danish (da)</td>
<td>Europarl</td>
<td>2,105</td>
<td>1,096</td>
</tr>
<tr>
<td>Dutch (nl)</td>
<td>Europarl</td>
<td>2,131</td>
<td>1,087</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>Europarl</td>
<td>2,038</td>
<td>1,049</td>
</tr>
<tr>
<td>Italian (it)</td>
<td>Europarl</td>
<td>2,135</td>
<td>1,087</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>Europarl</td>
<td>2,119</td>
<td>1,096</td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td>Europarl</td>
<td>2,153</td>
<td>1,107</td>
</tr>
<tr>
<td>Czech (cs)</td>
<td>ProjSynd</td>
<td>1,044</td>
<td>722</td>
</tr>
<tr>
<td>French (fr)</td>
<td>ProjSynd</td>
<td>885</td>
<td>657</td>
</tr>
<tr>
<td>German (de)</td>
<td>ProjSynd</td>
<td>1,086</td>
<td>769</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>ProjSynd</td>
<td>1,028</td>
<td>743</td>
</tr>
<tr>
<td>Albanian (sq)</td>
<td>SETimes</td>
<td>4,190</td>
<td>2,274</td>
</tr>
<tr>
<td>Bulgarian (bg)</td>
<td>SETimes</td>
<td>3,737</td>
<td>2,068</td>
</tr>
<tr>
<td>Croatian (hr)</td>
<td>SETimes</td>
<td>4,139</td>
<td>2,257</td>
</tr>
<tr>
<td>Greek (el)</td>
<td>SETimes</td>
<td>3,890</td>
<td>2,129</td>
</tr>
<tr>
<td>Macedonian (mk)</td>
<td>SETimes</td>
<td>3,573</td>
<td>1,956</td>
</tr>
<tr>
<td>Romanian (ro)</td>
<td>SETimes</td>
<td>4,355</td>
<td>2,368</td>
</tr>
<tr>
<td>Serbian (sr)</td>
<td>SETimes</td>
<td>3,943</td>
<td>2,156</td>
</tr>
<tr>
<td>Turkish (tr)</td>
<td>SETimes</td>
<td>4,040</td>
<td>2,196</td>
</tr>
<tr>
<td>Urdu (ur)</td>
<td>Urdu</td>
<td>1,828</td>
<td>1,093</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>55,206</td>
<td>29,533</td>
<td></td>
</tr>
</tbody>
</table>
Please identify Ahmet Sezer in the Turkish passage.

Ahmet Sezer

President Ahmet Sezer has accused the ruling party of trying to penetrate state administration with Islamic ideology.

In turn, Erdogan has criticised Sezer for blocking government appointments to public office.

Please identify Goran Kljajevic in the Turkish passage.

Goran Kljajevic

Among them are former Belgrade Commercial Court president Goran Kljajevic and a judge from that court, Delinka Djurdjevic.

Goran Kljajevic’s brother, Marko, was the head of the trial chamber in the Zoran Djindjic murder trial.

Marko Kljajevic withdrew from the trial in late August, objecting to the police and judiciary’s treatment of his brother.

5. Conclusion

We have demonstrated a methodology for creating and curating cross-language entity linking test collections, and used that methodology to create collections in twenty-one languages. We described how crowdsourced judgments can be used effectively to account for problems with bilingual projection of query names. Our approach uses existing aligned parallel corpora; this decision allows exploitation of existing high-quality English tools to economically obtain cross-language entity linking annotations. The collection is available at http://hltcoe.jhu.edu/datasets/.

6. 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, and Damianos Karakos, for which we are grateful. We are grateful to the creators of the Europarl (Koehn, 2005) and South-East European Times collections (Tyers and Alperen, 2010). Support for this work was provided, in part, by NSF grant CCF 0916081.

7. References


Chris Callison-Burch and Mark Dredze. 2010. 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


