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.[1] Very recently, TAC[2] 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

[1]http://ntcir.nii.ac.jp/CrossLink/
[2]http://nlp.cs.qc.cuny.edu/kbp/2011/

Table 1: 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 cumhurbaşkanlığı üyesi yarışını az farkla ikinci sırada tamamlayan **Haris Silaciç**, değişikliklere karşsı çıkıyor ve büyük bir destekçi kitlesine sahip bulunuyor.	E0305255	Haris Silajdži

Table 2: Our 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)

text collections offers the potential for further leverage, allowing the same ground truth English annotations to be projected to more than one language.

The parallel text collections we used are shown in Table 2. Together, these collections contain 196,717 non-English documents in five different scripts. To identify person names on the English side of each parallel text collection, we used the publicly available named entity recognition system created by Ratinov and Roth (2009); this resulted in 257,884 unique person name/document pairs across the six collections. We then eliminated all single-token names. Because named entity recognition is imperfect, we manually examined these English results to eliminate strings that were obviously not person names. We also eliminated names that occurred only once across the collection, and we limited to ten the number of times a single name string would be included (to avoid building a collection dominated by a small number of common names). We used person names exclusively in this collection; however, building test collections for other entity types, such as organizations, could be handled in the same way.

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 enough information." A single Mechanical Turk Human Intelligence Task (HIT) consisted of six such sets, two of which were interleaved queries for which we already knew ground truth. The three candidates were displayed in random order. We obtained three separate judgments for each query, and included the query 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. More details about this process appear in Mayfield et al. (2011).

3. Curation of Name Projections

Given an English entity linking query, name projection creates a corresponding cross-language entity linking query. We use a multi-step process to produce high quality (i.e., equivalent and error-free) names that have exact string matches in the non-English document. The following discussion uses Turkish as the canonical example, although all target languages followed a similar process.

The first step in our process uses the the Berkeley Word Aligner (Haghighi et al., 2009) to create a mapping from words in the English text to words or phrases in the Turkish parallel text. Second, for each English query name, a span of tokens in the Turkish document is associated with that name. This is based on the assumption that all names are written contiguously in the target language. This can compensate for the aligner missing the middle of a name, especially when the middle portion only appears in the Turkish document. It has the added benefit of making some misalignments obvious because of the large number of tokens included in the span. By aligning all names, rather than only those in the query set, the entire collection can be used to compensate for a misalignment in a particular document. The third step ranks all the projections for a single English name based on frequency. In the final step, the most frequent Turkish string appearing in the Turkish document is chosen as the projection for the English query name. Ties are broken based on the absolute difference in the number of tokens in the name between the two languages. A minimum difference is preferred, and having more tokens is preferred to having fewer tokens.

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

[3] http://translate.google.com/

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-mapping, which was used to estimate Turker accuracy. Each task was completed by three different individuals. When the Turkers agreed on a name projection, it was automatically accepted as the correct query string. Disagreements were resolved by an independent assessor familiar with the writing system. 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 goals of the collection and with the writing system but not the language 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.

Table 3: Language coverage in our collection.

Language	**Collection**	**Queries**	**Non-NIL**
Arabic (ar)	Arabic	2,829	661
Chinese (zh)	Chinese	1,958	956
Danish (da)	Europarl	2,105	1,096
Dutch (nl)	Europarl	2,131	1,087
Finnish (fi)	Europarl	2,038	1,049
Italian (it)	Europarl	2,135	1,087
Portuguese (pt)	Europarl	2,119	1,096
Swedish (sv)	Europarl	2,153	1,107
Czech (cs)	ProjSynd	1,044	722
French (fr)	ProjSynd	885	657
German (de)	ProjSynd	1,086	769
Spanish (es)	ProjSynd	1,028	743
Albanian (sq)	SETimes	4,190	2,274
Bulgarian (bg)	SETimes	3,737	2,068
Croatian (hr)	SETimes	4,139	2,257
Greek (el)	SETimes	3,890	2,129
Macedonian (mk)	SETimes	3,573	1,956
Romanian (ro)	SETimes	4,355	2,368
Serbian (sr)[5]	SETimes	3,943	2,156
Turkish (tr)	SETimes	4,040	2,196
Urdu (ur)	Urdu	1,828	1,093
Total		55,206	29,533

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

[4]This 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.

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

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 başkan 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 1: Example Turker Name Projection Tasks

Table 4: 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%
More descriptive name appears in document	1.1%
Manual name curation	5.0%
Only one occurrence of name in collection	15.8%
Ten occurrences of name already included	11.6%
Could not locate name in English document	0.5%
To avoid predicted NIL/non-NIL imbalance	4.0%

Table 5: Fraction of all queries lost during the human assessment phase.

Reason for Attrition	Queries Lost
Low Turker quality	0.9%
Turker disagreement	0.9%
Missing judgments	0.3%

across the 21 languages. Further attrition caused by projecting the English names onto those twenty-one languages, as shown in Table 5, resulted in a final non-English query count of 55,206.

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

Amazon.com. 2005. Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome.

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

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

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

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*

Mechanical Turk, CSLDAMT '10, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 923–931, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

James Mayfield, Dawn Lawrie, Paul McNamee, and Douglas W. Oard. 2011. Building a cross-language entity linking collection in 21 languages. In *Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation*.

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

Paul McNamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky, and Markus Dreyer. 2009. HLT-COE Approaches to Knowledge Base Population at TAC 2009. In *Proceedings of the Text Analysis Conference*.

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

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. 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. Association for Computational Linguistics.

Charu Tiwari, Pankaj Gulhane, Amit Madaan, and Rupesh Mehta. 2010. News assist: Identifying set of relevant entities used in news article. In *Forum for Information Retrieval Evaluation (FIRE)*.

Francis Tyers and Murat Serdar Alperen. 2010. 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.