HLTCOE Technical Reports

No. 6

Cross-lingual Coreference Resolution: A New Task for Multilingual Comparable Corpora

Spence Green, Nicholas Andrews, Matthew R. Gormley,
Mark Dredze and Christopher D. Manning

June 2011

Human Language Technology Center of Excellence
810 Wyman Park Drive
Baltimore, Maryland 21211
www.hltcoe.org

Spence Green
Nicholas Andrews
Matthew R. Gormley
Mark Dredze
Christopher D. Manning

Acknowledgment: This work is supported in part by the Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

Human Language Technology Center of Excellence
The Johns Hopkins University
810 Wyman Park Drive
Baltimore, Maryland 21211
410-516-4800
http://web.jhu.edu/HLTCOE/

Cross-lingual Coreference Resolution: A New Task for Multilingual Comparable Corpora

Spence Green[†*], Nicholas Andrews[†], Matthew R. Gormley[†], Mark Dredze[†], and Christopher D. Manning[*]

[†]Human Language Technology Center of Excellence, Johns Hopkins University
{noa,mrg,mdredze}@cs.jhu.edu

[*]Computer Science Department, Stanford University
{spenceg,manning}@stanford.edu

Abstract

We introduce *cross-lingual coreference resolution*, the task of grouping entity mentions with a common referent in a multilingual corpus. Information, especially on the web, is increasingly multilingual. We would like to track entity references across languages without machine translation, which is expensive and unavailable for many language pairs. Therefore, we develop a set of models that rely on decreasing levels of parallel resources: a bitext, a bilingual lexicon, and a parallel name list. We propose baselines, provide experimental results, and analyze sources of error. Across a range of metrics, we find that even our lowest resource model gives a 2.5% F1 absolute improvement over the strongest baseline. Our results present a positive outlook for cross-lingual coreference resolution even in low resource languages. We are releasing our cross-lingual annotations for the ACE2008 Arabic-English evaluation corpus.

1 Introduction

This work introduces *cross-lingual coreference resolution*, the task of clustering entity mentions across documents and languages. It consists of the following four components: (1) monolingual mention detection, (2) monolingual within-document coreference resolution, (3) cross-document clustering and (4) cross-lingual clustering. The first two are mention-level tasks performed for each document in a multilingual corpus. Together they are typically refered to as *coreference resolution* in the NLP community and comprise within-document anaphora resolution: the clustering of co-referent named entity, pronominal, and nominal mentions into chains. The final two are *entity-level* problems in which within-document coreference chains in multiple languages are clustered across documents. These are complementary tasks. For example, suppose five documents mention Qatar. The within-document system should produce one coreference chain per document with mentions such as "Qatar" or "2022 World Cup host." Then a cross-document system should merge the five chains into one entity corre-

Qatar

From Wikipedia, the free encyclopedia

Qatar (Arabic: قطر), also known as the State of Qatar or locally Dawlat Qatar, is an Arab country, known officially as an emirate, in the Middle East, occupying the small Qatar Peninsula on the northeasterly coast of the much larger Arabian Peninsula. It is bordered by Saudi Arabia...

(a) **Entities**: *Qatar, Wikipedia, Arab, Middle East, Qatar Peninsula, Arabian Peninsula, Saudi Arabia*

غيمة صناعية تبرد ملاعب كأس العالم في قطر

سي ان ان العربية

دبي، الإمارات العربية المتحدة قدم قسم الهندسة في جامعة قطر فكرة جديدة لمواجهة الحرارة المرتفعة في البلاد خلال مباريات كأس العالم التي فازت قطر بحق تنظيمها

(b) **Entities**: *World Cup stadium, Qatar, CNN Arabic, Eng. Division of Qatar University, World Cup, Qatar Organization*

Katar entwickelt Wolken-Drohne für WM

Kurier - Vor 19 Stunden

Um bei der Fußball-Weltmeisterschaft 2022 für ein wenig Schatten und damit Abkühlung in Stadien zu sorgen, wird in Katar an einer fliegenden Wolken-Drohne gearbeitet, berichtet die katarische Webseite "The Peninsula"...

(c) **Entities**: *Qatar, Wolken-Drohne, World Cup, Kurier, The Peninsula*

Figure 1: Entities (in order of appearance) in unaligned English, Arabic, and German documents. A cross-lingual coreference system would cluster the within-document coreference chains for Qatar (underlined).

sponding to the Arab Gulf state.

Previous work has investigated the within-document components for various languages. Therefore, we focus on the new cross-document cross-lingual components and develop models for cross-document, cross-lingual entity clustering.[1]

1.1 Task Example and Applications

As an example, consider references to Qatar in Figure 1. The Wikipedia entry for the country includes canonical English and Arabic spellings, two English aliases (one of which is a transliteration), and references to other geographic entities. The second and third articles (in Arabic and German) refer to the country's bid to host the 2022 World Cup. The orthographic forms "Qatar," قطر, and "Katar" all refer to the same real world entity, a conclusion that can be reached via mention similarity and context clues.

Applications for these models are abundant. For example, during the 2011 uprisings in the Middle East, emails, blog posts, Twitter messages, and other electronic texts were produced in English, French, various dialects of Arabic, and other languages. This type of data can be very difficult to automatically translate, yet we would still like to organize references to mentioned entities (e.g. "Wael Ghoneim", "Tahrir Square").

One research application of models for this task is entity linking, in which entity mentions are matched to corresponding entries in a knowledge base, such as Wikipedia (Bunescu and Pasca, 2006; Cucerzan, 2007). Present systems link entity mentions in English documents to English knowledge base entries (McNamee et al., 2010). But the world's information is increasingly multilingual, so the 2011 NIST Text Analysis Conference (TAC) task will consider cross-lingual linking. Another NLP application is machine translation (MT), in which name translation consistency (a particularly difficult MT problem (Hermjakob et al., 2008)) could be enforced by linking entity references in the source with a canonical spelling in the target.

In this work, we consider translating all documents to a common language—as has been done for related NLP tasks (Zitouni and Florian, 2008; Sayeed et al., 2009)—and running a mono-lingual cross-document coreference system. However, high quality statistical MT exists for only a fraction of the world's 7,000 natural languages (Lewis, 2009); Google Translate currently supports only 57. Also, names are especially problematic even for state-of-the-art MT systems. These two facts motivate us to develop models for settings without large bitexts. Our best low-resource model, which matches names with a simple classifier and maps mention contexts with a bilingual lexicon, achieves an 11.9% F1 absolute improvement over the cross-lingual baseline. We demonstrate that much can be achieved without the resources required for MT. [2]

2 Related Work

Since the introduction of the vector space model (VSM) (Bagga and Baldwin, 1998b), there has been comparatively little work on cross-document coreference resolution. In the VSM, mention contexts are mapped to feature vectors and clustered with a distance metric like cosine similarity. Successful feature extensions to the VSM have included biographical information (Mann and Yarowsky, 2003) and syntactic context (Chen and Martin, 2007). However, neither of these feature sets generalize easily to the cross-lingual setting with multiple entity types. Models simpler than the VSM have also been used. Mayfield et al. (2009) used a binary classifier to generate a directed graph of coreference chains, which were clustered by identifying connected components in the graph. Recent work has considered very large corpora (Rao et al., 2010; Singh et al., 2011).

Cross-document work on languages other than English is even more scarce. Wang (2005) used a combination of the VSM and heuristic feature selection strategies to cluster transliterated Chinese personal names. For Arabic, Magdy et al. (2007) started with the output of the mention detection and

[1] As a first step, we consider a language pair, but plan on extending to an arbitrary number of languages in future work.

[2] We assume that each document contains only one language. A useful extension to the task would be cross-lingual, within-document coreference resolution.

	Model 1	Model 2	Model 3
Context Mapping	MT	Lexicon	Polylingual Topic Model
Resource Requirements	Full MT resources (bitext, etc.)	Parallel name lists, bi-lingual lexicon	Parallel name lists, comparable corpora

Table 1: Cross-lingual coreference models based on natural extensions to existing cross-document coreference systems.

within-document coreference system of Florian et al. (2004). They clustered the entities incrementally using a binary SVM classifier. Baron and Freedman (2008) used complete-link agglomerative clustering, where merging decisions were based on a variety of features such as document topic and uniqueness of the name in Wikipedia. Finally, Sayeed et al. (2009) translated the Arabic name mentions to English and then formed clusters greedily using pairwise matching. This strategy is biased toward Arabic transliterations of English names.

To our knowledge, there has been no prior work on the cross-lingual task formulated in this paper. However, aspects of cross-lingual coreference resolution are addressed by related tasks that have overlapping definitions, objectives, and experimental settings:

- **Multilingual coreference resolution**: An umbrella term for the adaptation of monolingual within-document coreference models to languages other than English (Harabagiu and Maiorano, 2000; Luo and Zitouni, 2005).
- **Multilingual entity detection and tracking**: Same as multilingual (within-document) mention detection and coreference resolution, respectively (Florian et al., 2004).
- **Named entity translation**: For a non-English document, produce an inventory of entities in English. This was an ACE2007 pilot task (Song and Strassel, 2008).
- **Cross-language name search**: Match a *single* query name against a list of other multilingual names. Context is usually not considered (McCarley, 2009; Udupa and Khapra, 2010).
- **Cross-lingual coreference retrieval**: Same as cross-language name search, except focused on alias construction (Aktolga et al., 2008).
- **Cross-Document Person Name Resolution**: Distinguish between senses of the same name, e.g., whether "George Bush" refers to the 41st or 43rd American president (Fleischman and Hovy, 2004).

None of these tasks address the complete problem of clustering coreference chains across languages and documents. They do share common themes and components of our task. A practical benefit of our work is the consolidation and clarification of nomenclature and experimental designs.

3 Cross-lingual Models

We now present three models for cross-lingual coreference resolution, each of which contains four components. *Mention Matching* refers to determining equality between entity mention strings, which may be in different writing systems. *Context Mapping* refers to mapping mention contexts (i.e., the sentences containing a mention) to a common representation. *Context Matching* is the process of determining similarity between mapped contexts. Finally, *Constrained Clustering* is the grouping of coreference chains subject to linking constraints.

The three models differ only in terms of context mapping techniques (Table 1). For high resource language pairs, the most effective technique is to translate all mention contexts to the same language (in our case, English). For lower resource language pairs, we investigate deterministic mapping to English using a lexicon, and a mapping to 1-best topic assignments learned with a polylingual topic model (PLTM).

3.1 Mention Matching

We treat mention matching as a classification problem, which imposes constraints by determining which coreference chains cannot refer to the same entity based on their mentions. We use separate methods for within- and cross- language matching.

Jaro-Winkler (within language) When two mentions are in the same writing system, we use the Jaro-Winkler edit distance (Porter and Winkler, 1997), which Christen (2006) found to be a superior metric for name matching. Jaro-Winkler rewards matching prefixes, the empirical justification being that

less variation typically occurs at the beginning of names.[3] To use Jaro-Winkler for classification, we tune a threshold α on held-out data (we used $\alpha = 0.3$), where a score less than α indicates a match.

Log-linear (cross language) When mentions originate in different writing systems, edit distance calculations no longer apply. Our insight is that context-sensitive transport between orthographies—i.e., transliteration (Knight and Graehl, 1998)—is unnecessary so long as enough evidence exists to identify a match. We thus build a binary log-linear classifier that extracts features from aligned name pairs (Table 2). Prior to alignment, names are deterministically mapped to a common orthography.[4]

Since the mention strings are short, and the alignments are usually monotonic, we do not require full-blown word alignment. Instead, we treat alignment as a bipartite matching problem between strings e and f, where the edge weight between mapped tokens e_i and f_j is the Levenshtein edit distance. A minimum cost alignment can be found in cubic time with the Hungarian algorithm.[5]

From the aligned name pairs we can train a binary log-linear classifier using the feature functions defined in Table 2. We optimize the parameters $\boldsymbol{\lambda}$ with a gradient-based method that includes L_1 regularization (Andrew and Gao, 2007).

For the experiments in this paper, we train the classifier on automatic alignments, thus any parallel name list may be used as training data. Unlike prior work in coreference resolution, we do not skew the training data toward negative examples (since most names should not match). We improved accuracy by adding features that were predictive of the negative class. For example, we found that $\overline{\text{OVERLAP-}e}$ and $\overline{\text{OVERLAP-}f}$ significantly improved accuracy in the application setting.

OVERLAP-e, f	$\mathbb{1}(x)$ when $x \in \{e_i\} \bigcup \{f_j\}$ for $(i,j) \in a_{e,f}$
$\overline{\text{OVERLAP-}e}$	$\mathbb{1}(x)$ when $x \in \{e\},\ x \notin \{f\}$
$\overline{\text{OVERLAP-}f}$	$\mathbb{1}(x)$ when $x \in \{f\},\ x \notin \{e\}$
#BIGRAM-DIFFERENCE	Discretized value of abs($\lvert\{e\}\rvert - \lvert\{f\}\rvert$)
NORMALIZED-DICE	Discretized value of $\frac{1}{n}\sum_{(i,j)\in a_{e,f}} \text{Dice}(e_i, f_j)$, where $n = \max(\lvert e\rvert, \lvert f\rvert)$
#"TRANSLATIONS"	Discretized number of aligned tokens with $(i,j) \in a_{e,f}$ and Lev$(e_i, f_j) > 3.0$
EDIT-DISTANCE	Discretized value of $\sum_{(i,j)\in a_{e,f}} \text{Lev}(e_i, f_j)$
LENGTH-SYMMETRY	$\mathbb{1}(\lvert e\rvert > \lvert f\rvert)$ or $\mathbb{1}(\lvert e\rvert < \lvert f\rvert)$ or $\mathbb{1}(\lvert e\rvert = \lvert f\rvert)$
LENGTH-DIFFERENCE	Discretized value of abs($\lvert e\rvert - \lvert f\rvert$)
IS-SINGLETON-PAIR	$\mathbb{1}(\lvert e\rvert = 1 \wedge \lvert f\rvert = 1)$
e-SINGLETON	$\mathbb{1}(\lvert e\rvert = 1)$
f-SINGLETON	$\mathbb{1}(\lvert f\rvert = 1)$

Table 2: Language-independent feature templates for a name pair $\langle e, f\rangle$ with alignment $a_{e,f}$. $\{\cdot\}$ indicates the collection of bigrams in a string. $|\cdot|$ is the (whitespace delimited) token length of a string. Lev$(\cdot,\cdot)$ is the Levenshtein edit distance between two strings. Among the most predictive features are OVERLAP-e, f bigrams at the beginning of aligned tokens. This is the same intuition behind Jaro-Winkler.

3.2 Context Mapping

If two mentions match, then we use context for further disambiguation. As with the mentions, the contexts may originate in different writing systems. We present two approaches to mention context mapping which differ in the resources required for the language pair (a lexicon versus comparable corpora).

Machine Translation For our evaluation, we translated all documents to English using the MT system Phrasal (Cer et al., 2010) which, like most public MT systems, lacks a transliteration module.[6] We believe that this approach yields the strongest expected results as it relies on years of resource cultivation and MT system development. We refer to this approach as Model 1.

Lexicon A simple deterministic approach is to map all contexts to a common language using a bilingual lexicon. For each context, we greedily map spans of words that appear in a given lexicon and drop all other words. We refer to this as Model 2.

Polylingual Topic Model Consistent with our emphasis on settings that lack parallel training data,

[3] For multi-token names, we sort the tokens prior to computing the score.

[4] This idea is reminiscent of Soundex, which Freeman et al. (2006) used for cross-lingual name matching.

[5] Names are characteristically short, and we do not observe a prohibitive run-time penalty from executing an $O(n^3)$ algorithm during clustering.

[6] §A.1 contains the complete MT system configuration.

we consider a largely unsupervised mapping. The polylingual topic model (PLTM) (Mimno et al., 2009) is a generative process in which document tuples (with one document per language) share a topic distribution. However, our setting assumes *unaligned* multilingual corpora. To increase vocabulary coverage for such a setting, Mimno et al. (2009) suggested construction of a corpus consisting of topically-aligned tuples in addition to singleton in-domain documents. The topically-aligned tuples serve as "glue" to share topics between languages, while the in-domain documents distribute those topics over in-domain vocabulary.[7] In total, we estimate two topic-word distributions (one for each language), and a single document-topic distribution linking the two languages. At test time, we infer the 1-best topic for context words which we use as an abstract representation of the observed words and provide the clustering algorithm with these topics in place of words. We refer to this as Model 3.

3.3 Context Matching

Once the contexts are mapped, we can compute a similarity score. Since each model maps words to a shared space (either all English or topics), we can represent mention contexts using entity language models (ELM) (Raghavan et al., 2004). Consider a set of coreference chains $d \in E_i$, which represents an entity. The context for each mention $m \in d_j$ is the sentence containing m. We concatenate the contexts to form a multiset of words S_i from which we can estimate a unigram ELM. For each word $w \in S_i$, we use a parameter estimate that includes a unigram prior P_c estimated from the entire corpus:

$$P_{E_i}(w|S_i) = \frac{\text{count}_w(S_i) + \beta P_c(w)}{|S_i| + \beta}$$

where we tune β on a development set. Given entity E_k (context S_k), we compare ELMs using the square root of twice the Jensen-Shannon divergence:

$$\begin{aligned} sim(P_{E_i}, P_{E_k}) &= \sqrt{2 \cdot JSD(P_{E_i}||P_{E_k})} \\ &= \sqrt{KL(P_{E_i}||M) + KL(M||P_{E_k})} \end{aligned}$$

where $KL(P_{E_i}||M)$ is the Kullback-Leibler divergence and $M = \frac{1}{2}(P_{E_i} + P_{E_k})$ (Endres and Schindelin, 2003).

[7] Mimno et al. (2009) showed that so long as the proportion of topically-aligned to non-aligned documents exceeded 0.25, the topic distributions (as measured by mean Jensen-Shannon divergence between distributions) did not degrade significantly.

3.4 Constrained Clustering

All three models use group-average hierarchical agglomerative clustering (HAC) in which each cluster has an ELM estimated from all contexts of all mentions in the cluster. We re-estimate the cluster ELM at each merge step. There are four possible parameters, which we tune on a development set: α, the Jaro-Winkler cutoff; β, weight of corpus prior in each ELM; δ, the distance at which clustering is terminated; and, k, the number of PLTM topics.

3.4.1 Clustering Constraints

Our coreference model also encodes entity-level constraints.[8] Prior to clustering, we remove all links in the proximity matrix that violate several pairwise binding constraints.[9] Consequently, context is only used to disambiguate the mention string confusable entities. Empirically, this strategy both improves performance and reduces runtime. The merging of coreference chains d_i and d_j is *disallowed* if:

1. **Document origin**: $doc(d_i) = doc(d_j)$ Do not merge chains from the same document since we assume prior within-document disambiguation.
2. **Semantic type**: $type(d_i) \neq type(d_j)$ Do not merge chains with different semantic types.
3. **Mention Match**: $f(m_i, m_j) = false$, where $m_i \in mentions(d_i)$ and $m_j \in mentions(d_j)$, and $f(\cdot)$ is the mention matching method from §3.1.

The $mentions(\cdot)$ function returns the *representative mention* of each entity, which is the first mention of that entity in a document. In many languages, the first mention is typically more complete than later mentions. Crucially, this heuristic largely decouples our model from within-document systems.[10]

In the cross-lingual case, we emphasize that $f(\cdot)$ is more likely to produce false positives than exact

[8] Albeit rather fewer than within-document models since syntactic constraints are largely irrelevant.

[9] Constraints like ours can be interpreted as high recall heuristics, which have been variously called *pair-filters* (Mayfield et al., 2009) and *sieves* (Raghunathan et al., 2010).

[10] Although in this work we use *all* mention contexts to estimate ELMs. Our model could work with the representative mentions only. In that case, expanding the mention context could combat the potential sparsity issue.

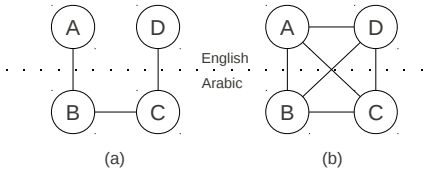

Figure 2: Cross-lingual transitivity. An arc between two entities indicates that they can link. (a) Transitivity is enforced in an ad-hoc manner. In this case, A and D could be linked through the Arabic. (b) Transitive closure is explicit: entities must be a fully-connected component.

matching. Ideally, spelling variation of transliterated names is a problem that we could partially resolve with knowledge of the spelling in the source language. We could thus benefit from ad-hoc transitivity by linking entities in one language through entities in the source language (Figure 2(a)). However, we find that for our current models and corpora, the transitive closure constraint in Figure 2(b) (Finkel and Manning, 2008), gives better results.

4 Training Data

Models 2 and 3 do not require high resource parallel bitexts. However, we evaluate our method on Arabic-English, a high resource language pair, since the only corpus that supports evaluation—the ACE2008 evaluation set—contains documents in these languages, and we can train Model 1 for comparison. In addition to the evaluation corpus, we used several other data resources, all of which could be cheaply obtained for many languages.

Orthographic Mapping The mention matching algorithm requires a common orthography for names. We use a simple, deterministic mapping from Arabic to the Latin alphabet (Table 8 in §8). A key feature of our mapping is that we remove most vowels. Because Arabic text is conventionally unvocalized, the diacritized short vowels, which are dropped, are inconsistently transliterated. Rather than trying to infer the diacritics, we eliminate all but one of the English vowels. Mappings for other language pairs could be quickly designed, or machine transliteration systems could be used in place of our log-linear model. For example, Irvine et al. (2010) show how to create machine transliterators cheaply for 150 languages using Wikipedia.

Parallel Name List The log-linear mention matching model is trained on a parallel name list mapped with Table 8. Name pair lists can be obtained from the LDC (e.g., LDC2005T34 contains nearly 450,000 parallel Chinese-English names) or Wikipedia (more than 200 languages). We extracted 12,860 name pairs from the parallel Arabic-English translation treebanks,[11] although we show in §6.1 that significantly fewer names are actually required. We generate a uniform distribution of training examples by flipping a coin for each aligned name pair in the corpus. If the coin is heads, we replace the English name with another English name chosen randomly from the corpus. We emphasize that the word-to-word alignments in our training data are not used in these experiments.

Bilingual Lexicon Lexicons are available or can be bootstrapped for numerous language pairs (Haghighi et al., 2008; Irvine and Klementiev, 2010). We compiled a lexicon from the English-Arabic translation treebanks, OmegaWiki, and the Universal Dictionary. Lexicon conflicts were resolved by voting. The composite lexicon contains 31,307 Arabic entries including 4,452 multiword entries. However, our analysis in §7 shows that high accuracy can be achieved with fewer than 1,000 entries.

PLTM Document Tuples Cross-lingual links in Wikipedia are abundant: as of February 2010, there were 77.07M cross-lingual links among Wikipedia's 272 language editions (de Melo and Weikum, 2010). For PLTM training, we formed a corpus of 19,139 English-Arabic topically-aligned Wikipedia "glue" articles, and 20,000 document singletons from the ACE2008 training corpus.

5 Evaluation Framework

Our experimental design is a cross-lingual extension of the Automatic Content Extraction (ACE) 2008 cross-document task (Strassel et al., 2008; NIST, 2008). We evaluate on name (NAM) mentions for cross-lingual person (PER) and organization (ORG) entities. We presume neither the number of entities

[11] LDC Catalog numbers LDC2009E82 and LDC2009E88. Since training ignores the word-to-word alignments, any corpus of parallel names will suffice.

nor their attributes (i.e., the task does not include a knowledge base).

A comprehensive cross-lingual coreference evaluation would include monolingual mention detection and within-document resolution. However, the task can be factorized, and the main contribution of this paper is cross-lingual, cross-document clustering. As a result, our experiments assume gold monolingual decisions: mention boundaries, semantic types, and within-document coreference chains.

Gold Coreference Chains A SemEval 2010 shared task showed that monolingual mention detection and within-document coreference accuracy can vary widely (Recasens et al., 2010). This variability makes the cross-lingual entity-based evaluation less interpretable. Further, Haghighi and Klein (2010) show that confusion between PER and ORG entities is not a significant source of error in state-of-the-art within-document coreference systems. Our experiments are consistent with prior work on within-document coreference resolution in which previous pipeline stages were assumed to produce gold output in order to isolate later stages (Bengtson and Roth, 2008).

Entity Level Evaluation We perform an *entity-level evaluation*, i.e., the coreference chain is the unit of evaluation, not each mention, so the use of gold mention chains does not inflate accuracy. Our models use the first mention string of each entity in a document, so the presence of additional mentions merely provides extra context for disambiguation.

Coreference evaluations are particularly sensitive to corpus configurations, annotation schemes, and shortcomings of the clustering metrics (Recasens and Hovy, 2010). To better quantify model performance, it has become customary to report multiple metrics. We validate our results with four different *entity-based* metrics:

- B^3 (Bagga and Baldwin, 1998a): Precision and recall are computed from the intersection of the hypothesis and reference clusters.
- CEAF (Luo, 2005): Precision and recall are computed from a maximum bipartite matching between hypothesis and reference clusters.
- VI and NVI (Reichart and Rappoport, 2009): Information-theoretic measures that utilize the entropy of the clusters and their mutual information. Unlike VI, normalized VI (NVI) is not sensitive to the size of the data set.

	Docs	Words	Tokens	Entities	Single	Mentions
Arabic	412	32,822	178,269	4,216	2,168	9,222
English	414	18,964	246,309	3,950	1,826	9,140

Table 3: ACE2008 PER and ORG entity statistics. The presence of singletons is a significant difference between ACE and MUC corpora.

5.1 Evaluation Corpus

The automatic evaluation of cross-lingual coreference systems requires annotated multilingual corpora. Monolingual cross-document annotation is expensive (Strassel et al., 2008), so we chose the ACE2008 Arabic-English evaluation corpus as a starting point for cross-lingual annotation. The corpus consists of 412 Arabic and 414 English unaligned documents sampled from independent sources over the course of a decade from seven genres (Table 3). Monolingual cross-document coreference linking is provided for 8,166 PER and ORG entities, which together have 18,362 NAM mentions.[12] From these, we found and annotated 216 cross-lingual entities.[13] To our knowledge, this is the first cross-lingual coreference annotated corpus.

6 Experiments

6.1 Mention Matching

We evaluated the cross-lingual mention matching classifier independently of the coreference model (Table 4) using a random 80/10/10 (train, development, test) split of the corpus. Of the mis-classified examples, we observed three major error types. First, the model learns that high edit distance is predictive of a mis-match. However, singleton strings that do not match often have a lower edit distance than longer strings that do match. As a result, singletons often cause false positives. Second, names that originate in a third language tend to violate the phonemic correspondences. For example, the model gives a false negative for a German football team: اف سي كايزرسلاوترن (*af s kazrslawtrn* using mapping) versus FC Kaiserslautern (*fc kasrslatrn*). Finally, translations, less common for personal names, are prob-

[12] §A.2 describes particulars of the data preparation.

[13] The annotators were the first author and another non-native speaker of Arabic. The annotations, corrections, and corpus split are available at http://www.cs.jhu.edu/~mdredze.

	Genre	Train	Test	Acc. (%)
JARO-WINKLER	all		1286	89.5
LOG-LINEAR	all	10,288	1286	**97.1 (+7.55)**
	nw	7,443	930	96.6
	bn	2,720	340	95.6
	wb	125	16	87.5

Table 4: Cross-lingual name matching accuracy [%]. We trained the binary classifier (LOG-LINEAR) on each name pairs from each genre separately (bn = broadcast news; nw = newswire; wb = weblog). As a baseline, we ran Jaro-Winkler on the mapped representation of each name pair (Table 8). Although we use the full training corpus for coreference experiments, high accuracy was possible with significantly fewer examples (bn).

	CEAF↑	VI↓	NVI↓	B^3 ↑			
				#hyp	P	R	F1
Monolingual Arabic (#gold=1,721)							
MODEL 0	86.0	0.419	0.060	1,873	93.7	83.7	88.4
+ CONTEXT	**87.2**	**0.368**	**0.052**	1,669	89.8	89.8	**89.8**
Monolingual English (#gold=1,529)							
MODEL 0	86.4	0.379	0.056	1,801	98.6	80.5	88.6
+ CONTEXT	**88.5**	**0.282**	**0.0420**	1,536	93.7	89.0	**91.4**

Table 5: Monolingual cross-document coreference evaluation (no cross-lingual linking attempted).

lematic. For example, the classifier produces a false negative for ⟨God, *gd*⟩ $\stackrel{?}{=}$ ⟨الله, *allh*⟩.

6.2 Cross-lingual Coreference Experiments

We evaluate all models using the experimental design from §5. In addition to B^3, CEAF, VI, and NVI, we include a separate evaluation of the cross-lingual (target) entities. We propose a modification of B^3 called B^3_{target}; only target entities are evaluated, with spurious non-target entities in the clustering solution penalized by a parameter (ranging between 0 to 1.) The relative ranking of our results remains consistent for different values of the penalization parameter (see §A.4).

For comparison, we first provide a standard monolingual cross-document coreference evaluation for each part of the corpus (Arabic and English) (Table 5). The baseline, MODEL 0, is very similar to our other models (constrained hierarchical clustering, Jaro-Winkler mention matching) only without cross-lingual context mapping. We give results both with and without context disambiguation.

Table 6 contains results of the cross-lingual evaluation, the main contribution of this paper. We provide two cross-lingual baselines:

- NAIVE: Each cluster has only one coreference chain. The high proportion of singletons in ACE corpora can inflate evaluation metrics (Table 3).
- CONSTRAINTS: Cluster are fully-connected components subject to entity-level constraints (§3.4.1).

7 Discussion

Across a range of metrics, we find that both our high and low resource models significantly improve over the cross-lingual baselines. Due to the presence of singletons, NAIVE does well on B^3 and CEAF. However, since cross-lingual entities by definition contain at least two coreference chains, CONSTRAINTS exceeds NAIVE on B^3_{target}.

The ordering of Models 1-3 corresponds to the degree of required parallel training resources. Model 1 (MT) performs in the range of within-language coreference models, suggesting that MT is an effective approach to cross-lingual coreference. Although Model 3 (PLTM) lags MT, it still produces a significant improvement over both cross-lingual baselines with minimal resources. Additionally, Model 2 achieves half of the accuracy improvement of Model 1 with significantly fewer resources, a good sign for low resource languages. High frequency ORG entities are the major source of error for Models 2-3, which suggests that focusing on these entities could further improve the low resource models. We show some system output in Table 7.

The PLTM, an unsupervised method, achieves performance comparable to Model 2 across several metrics. It is thus a viable alternative when a lexicon either does not exist for a particular language pair or lacks domain-specific terms that are necessary to disambiguate entities with similar names. Another advantage of the PLTM is that resources are not required for each language pair: for n languages, the training corpus consists of tuples of n documents about the same topics, one in each language. It would be easy to obtain these documents from Wikipedia cross-lingual links, and this suggests a direction for future work on many simultaneous languages.

Our best low resource model uses a large lexicon.

	CEAF↑	VI↓	NVI↓	B^3 ↑					B^3_{target} ↑				
				#gold	#hyp	P	R	F1	#gold	#hyp	P	R	F1
English-Arabic — Baselines													
NAIVE	**64.9**	1.22	0.165	3,057	5,453	100.0	56.1	**71.8**	146	1,587	100.0	9.20	16.9
CONSTRAINTS	57.4	**1.00**	**0.136**	3,057	2,216	65.6	75.2	70.1	146	517	78.3	41.8	**54.5**
English-Arabic — This Paper													
MODEL 1 (MT)	**80.3**	**0.512**	**0.069**	3,057	2,783	85.4	85.8	**85.6**	146	310	93.2	67.7	**78.4**
MODEL 2 (Lexicon)	73.3	0.687	0.093	3,057	2,610	78.8	82.0	80.4	146	395	87.6	53.5	66.4
MODEL 3 (PLTM)	72.1	0.810	0.110	3,057	2,746	77.5	77.2	77.3	146	506	84.2	43.1	57.0

Table 6: Cross-lingual coreference evaluation. Higher scores (↑) are better for CEAF and B^3, whereas lower (↓) is better for VI and NVI. #gold indicates the number of *unique* entity ids, whereas #hyp is the number of clusters produced by each system. B^3_{target} scores cross-lingual entities only with a non-target entity weight of 1.0. For B^3_{target}, cross-lingual model performance is correlated with the degree of parallel training resources.

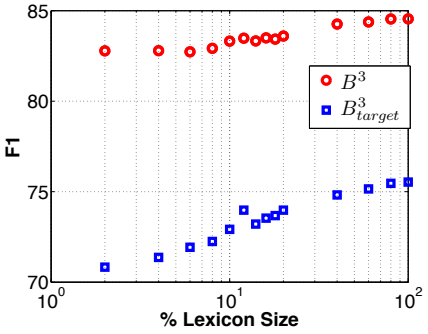

Figure 3: Model 2 for various lexicon sizes (development set). Model 0 + Context achieves 73.41 B^3 and 60.45 B^3_{target}, which Model 2 exceeds with 2% of its lexicon.

To quantify how many entries are actually required, we filtered the Arabic lexicon based on development set word frequency. Figure 3 shows that B^3_{target} accuracy decreases by only 4.70 F1 when just 2% (626) of the original entries remain. We found that high frequency words accounted for most of the overlap between any two ELMs, hence the minimal degradation in performance when low frequency words were removed.

8 Conclusion

We have introduced cross-lingual coreference resolution, the first annotations for this task, and models for low and high resource settings. Future experimental work will include quantifying the impact of errorful within-document chains, jointly inducing within- and cross-document coreference clusters, and extending to many languages. The corpus we release, and the experimental procedure we have specified, will facilitate these developments.

Acknowledgments We thank Angel Chang, Nate Chambers, Ken Church, Jason Eisner, David Mimno, Scott Miller, Jim Mayfield, Paul McNamee, and Val Spitkovsky for helpful discussions.

Merged Entities					
1	Hamed bin Khalifa Al-Thani	حمد بن خليفة آل ثاني	*hmd bn khlfah al than*	Referent	Transliterate
2	Arab League	الجامعة الدول العربية	*aljamaah aldwl alarbah*	Referent	Translate
3	Venezuela General Oil Company	منظمة الدول المصدرة للنفط	OPEC	¬ Referent	Translate
4	Abdullah bin Hussein	عبد الله بن عبد العزيز	*abd allh bn abd alazz*	¬ Referent	Transliterate
Non-merged Entities					
5	Agence France Press	وكالة الانباء الفرنسية	*wkalah alanba alfrnsah*	Referent	Translate
6	CIA	وكالة الاستخبارات المركزية	*wkalah alastkhbarat almrkzah*	Referent	Translate

Table 7: Entities from Model 2's clustering (dev set). In *italics* we give either a mapping (Table 8) or a translation of the Arabic. (1-4) all pass the approximate mention match constraint. It is likely that the entities in (3) have similar contexts, complicating cross-lingual disambiguation. (4) is harder, as the current monarchs of Jordan and Saudi Arabia have similar names and contexts. (5-6) are high frequency ORG entities present in the MT phrase table but not the lexicon. Evaluated on PER entities only, Model 2 obtains 80.03 F1 v. 83.10 F1 for Model 2 (B^3_{target}; development set).

References

E. Aktolga, M. Cartright, and J. Allan. 2008. Cross-document cross-lingual coreference retrieval. In *CIKM*.

G. Andrew and J. Gao. 2007. Scalable training of L1-regularized log-linear models. In *ICML*.

A. Bagga and B. Baldwin. 1998a. Algorithms for scoring coreference chains. In *LREC*.

A. Bagga and B. Baldwin. 1998b. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*.

A. Baron and M. Freedman. 2008. Who is Who and What is What: Experiments in cross-document co-reference. In *EMNLP*.

E. Bengtson and D. Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.

R. C. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

D. Cer, M. Galley, D. Jurafsky, and C. D. Manning. 2010. Phrasal: A statistical machine translation toolkit for exploring new model features. In *HLT-NAACL, Demonstration Session*.

Y. Chen and J. Martin. 2007. Towards robust unsupervised personal name disambiguation. In *EMNLP-CoNLL*.

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

S. Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP*.

G. de Melo and G. Weikum. 2010. Untangling the cross-lingual link structure of Wikipedia. In *ACL*.

D. M. Endres and J. E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858 – 1860.

J. R. Finkel and C. D. Manning. 2008. Enforcing transitivity in coreference resolution. In *ACL-HLT*.

M. Fleischman and E. Hovy. 2004. Multi-document person name resolution. In *ACL Workshop on Reference Resolution and its Applications*.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, et al. 2004. A statistical model for multilingual entity detection and tracking. In *HLT-NAACL*.

A. T. Freeman, S. L. Condon, and C. M. Ackerman. 2006. Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance algorithm. In *HLT-NAACL*.

M. Galley and C. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP*.

M. Galley, S. Green, D. Cer, P. C. Chang, and C. D. Manning. 2009. Stanford University's Arabic-to-English statistical machine translation system for the 2009 NIST evaluation. Technical report, Stanford University.

S. Green, M. Galley, and C. D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *HLT-NAACL*.

N. Habash and O. Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL*.

A. Haghighi and D. Klein. 2010. Coreference resolution in a modular, entity-centered model. In *NAACL*.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*.

S. M. Harabagiu and S. J. Maiorano. 2000. Multilingual coreference resolution. In *ANLP*.

U. Hermjakob, K. Knight, and H. Daumé. 2008. Name translation in statistical machine translation: Learning when to transliterate. In *ACL*.

A. Irvine and A. Klementiev. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proc. of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

A. Irvine, C. Callison-Burch, and A. Klementiev. 2010. Transliterating from all languages. In *AMTA*.

K. Knight and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24:599–612.

M. P. Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International.

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *NAACL*.

X. Luo and I. Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In *HLT-EMNLP*.

X. Luo. 2005. On coreference resolution performance metrics. In *HLT-EMNLP*.

W. Magdy, K. Darwish, O. Emam, and H. Hassan. 2007. Arabic cross-document person name normalization. In *Workshop on Computational Approaches to Semitic Languages*.

G. S. Mann and D. Yarowsky. 2003. Unsupervised personal name disambiguation. In *NAACL*.

J. Mayfield, D. Alexander, B. Dorr, J. Eisner, T. Elsayed, et al. 2009. Cross-document coreference resolution: A key technology for learning by reading. In *AAAI Spring Symposium on Learning by Reading and Learning to Read*.

J. S. McCarley. 2009. Cross language name matching. In *SIGIR*.

P. McNamee, H. T. Dang, H. Simpson, P. Schone, and S. M. Strassel. 2010. An evaluation of technologies for knowledge base population. In *LREC*.

D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. In *EMNLP*.

NIST. 2008. Automatic Content Extraction 2008 evaluation plan (ACE2008): Assessment of detection and recognition of entities and relations within and across documents. Technical Report rev. 1.2d, National Institute of Standards and Technology (NIST), 8 August.

E. H. Porter and W. E. Winkler, 1997. *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, chapter 6, pages 190–199. U.S. Bureau of the Census.

H. Raghavan, J. Allan, and A. McCallum. 2004. An exploration of entity models, collective classification and relation description. In *KDD Workshop on Link Analysis and Group Detection*.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M Surdeanu, D. Jurafsky, and C. D. Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*.

D. Rao, P. McNamee, and M. Dredze. 2010. Streaming cross document entity coreference resolution. In *COLING*.

M. Recasens and E. Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *ACL*.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, et al. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

R. Reichart and A. Rappoport. 2009. The NVI clustering evaluation measure. In *CoNLL*.

A. Sayeed, T. Elsayed, N. Garera, D. Alexander, T. Xu, et al. 2009. Arabic cross-document coreference detection. In *ACL-IJCNLP, Short Papers*.

S. Singh, A. Subramanya, F. Pereira, and A. McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *ACL*.

Z. Song and S. Strassel. 2008. Entity translation and alignment in the ACE-07 ET task. In *LREC*.

S. Strassel, M. Przybocki, K. Peterson, Z. Song, and K. Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *LREC*.

R. Udupa and M. M. Khapra. 2010. Improving the multilingual user experience of Wikipedia using cross-language name search. In *HLT-NAACL*.

H. Wang. 2005. Cross-document transliterated personal name coreference resolution. In L. Wang and Y. Jin, editors, *Fuzzy Systems and Knowledge Discovery*, volume 3614 of *Lecture Notes in Computer Science*, pages 11–20. Springer.

I. Zitouni and R. Florian. 2008. Mention detection crossing the language barrier. In *EMNLP*.

A Appendix

A.1 MT System Description

MODEL 1 uses MT for context mapping. The MT system is Phrasal (Cer et al., 2010) with the Moses baseline feature set except for linear distortion, to which we added future cost estimation (Green et al., 2010). We also included the hierarchical lexicalized re-ordering models of Galley and Manning (2008). To tune parameters, we ran MERT with the Downhill Simplex algorithm on the MT06 dataset.[14]

The training corpus was all data permitted under the NIST OpenMT 2009 constrained track evaluation. We created word alignments using the Berkeley Aligner (Liang et al., 2006) and symmetrize using the grow-diag heuristic. We built a 5-gram language model from the Xinhua and AFP sections of the Gigaword corpus (LDC2007T07), in addition to all of the target side training data. The language model was smoothed with the modified Kneser-Ney algorithm.

A.2 Data Preprocessing

Prior to the ACE2008 evaluation, LDC only performed annotation and quality control for 50 monolingual target entities. Subsequent to the evaluation, LDC added cross-document annotation for the remaining PER and ORG entities, but was unable to complete quality checking (*p.c.*). For example, the

[14] This is the same Ar-En baseline configuration as (Galley et al., 2009), which placed 2nd in the NIST 2009 OpenMT evaluation. For comparison, the system achieved 53.33 BLEU-4 on MT03. The training data is available at http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

Arabic Rules			
ب → b	ت → t	ث → th	ج → j
ح → h	خ → kh	د → d	ذ → th
ر → r	ز → z	س → s	ش → sh
ص → s	ض → d	ط → t	ظ → th
ع → a	غ → g	ف → f	ق → q
ك → k	ل → l	م → m	ن → n
ه → h	ا → a	و → w	ى → a
ة → ah		ء, ي →	ئ, ؤ, - →
English Rules			
k → c	**p** → b	**x** → ks	**e,i,o,u** →

Table 8: Deterministic English-Arabic orthographic mapping to a common orthographic representation.

entity القاعدة *Al-Qaeda* has at least six different cross-document ids. We corrected these errors in the cross-lingual annotation. We also excluded the English document ABC19980519.1830.0856.LDC2000T44, which has incorrect character offsets.

Both the ACE2008 and Wikipedia documents are unprocessed. For English, we tokenized and split sentences using packages from the Stanford parser, and stemmed tokens using the Porter algorithm. For Arabic, we removed diacritics, applied simple orthographic normalization, and segmented clitics with MADA (Habash and Rambow, 2005).

Because a similar corpus did not exist for development, we split the ACE2008 evaluation corpus. However, the usual method of splitting at the document level would not ensure that all mentions of a given entity were confined to one side of the split. We thus split the corpus *by entity id*. Since some cross-lingual entities occur disproportionately in the corpus (e.g. "Xinhua", "Agence France Presse"), we created partitions using frequency-matched stratified sampling. We assigned one third of the target cross-lingual entities to development, and the remaining target entities to test. We partitioned the non-target mono-lingual entities similarly.

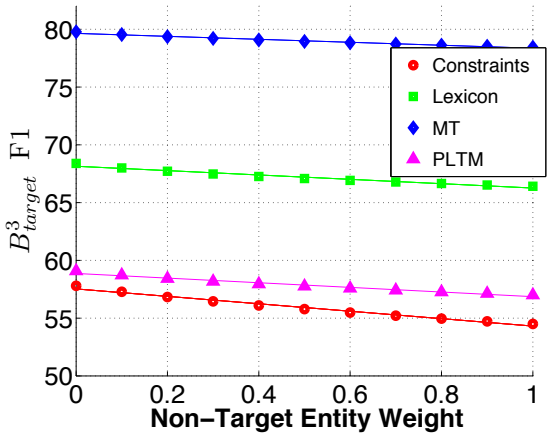

Figure 4: B^3_{target} (test set; shown with least squares regression lines) is sensitive to the presence of non-target entities, but the ranking of methods does not change across a range of non-target entity weights.

A.3 English-Arabic Orthographic Mapping

To train the log-linear mention matching classifier, we converted the training data to a common orthography using the mapping in Table 8.

A.4 Description of B^3_{target} Metric

Scoring a subset of a clustering solution requires caution. Consider B^3 precision for target entities T, where $t_i \in T$ has hypothesis cluster H_{t_i} and gold cluster G_{t_i}:

$$P = \frac{1}{|T|} \sum_{t_i}^{T} \frac{|H_{t_i} \bigcap G_{t_i}|}{|H_{t_i}|} \quad (1)$$

The presence of spurious non-target entities in H_{t_i} affects the denominator of (1). Removing those entities would inflate precision, while retaining them could result in an excessive penalty, which would make the cross-lingual metric less informative. A compromise solution is to weight non-target entity contributions to (1), where lower weights discount possible penalties. We call the modified metric B^3_{target}, and consider various weights in Figure 4. It is unclear how to change CEAF and NVI similarly. Strassel et al. (2008) discuss the issue of spurious entity mentions, but do not indicate whether NIST included them in the final evaluation.