# Entity Clustering Across Languages

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

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

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

## Abstract

Standard entity clustering systems commonly rely on mention (string) matching, syntactic features, and linguistic resources like English WordNet. When co-referent text mentions appear in different languages, these techniques cannot be easily applied. Consequently, we develop new methods for clustering text mentions across documents and languages simultaneously, producing cross-lingual entity clusters. Our approach extends standard clustering algorithms with cross-lingual mention and context similarity measures. Crucially, we do not assume a pre-existing entity list (knowledge base), so entity characteristics are unknown. On an Arabic-English corpus that contains seven different text genres, our best model yields a 24.3% F1 gain over the baseline.

## 1 Introduction

This paper introduces techniques for clustering coreferent text mentions across documents and languages. On the web today, a breaking news item may instantly result in mentions to a real-world entity in multiple text formats: news articles, blog posts, tweets, etc. Much NLP work has focused on model adaptation to these diverse text genres. However, the *diversity of languages* in which the mentions appear is a more significant challenge. This was particularly evident during the 2011 popular uprisings in the Arab world, in which electronic media played a prominent role. A key issue for the outside world was the aggregation of information that appeared simultaneously in English, French, and various Arabic dialects.

To our knowledge, we are the first to consider clustering entity mentions across languages without *a priori* knowledge of the quantity or types of real-world entities (a knowledge base). The cross-lingual setting introduces several challenges. First, we cannot assume a prototypical name format. For example, the Anglo-centric first/middle/last prototype used in previous name modeling work (*cf.* (Charniak, 2001)) does not apply to Arabic names like *Abdullah ibn Abd Al-Aziz Al-Saud* or Chinese names like *Hu Jintao* (referred to as *Mr. Hu*, not *Mr. Jintao*). Second, organization names often require both transliteration and translation. For example, the Arabic شركة جنرل موتورز 'General Motors Corp' contains transliterations of جنرل موتورز 'General Motors', but a translation of شركة 'Corporation'.

Our models are organized as a pipeline. First, for each document, we perform standard mention detection and coreference resolution. Then, we use pairwise cross-lingual similarity models to measure both mention and context similarity. Finally, we cluster the mentions based on similarity.

Our work makes the following contributions: (1) introduction of the task, (2) novel models for cross-lingual entity clustering of person and organization entities, (3) cross-lingual annotation of the NIST Automatic Content Extraction (ACE) 2008 Arabic-English evaluation set, and (4) experimental results using both gold and automatic within-document processing. We will release our software and annotations to support future research.

### 1.1 Task Description via a Simple Example

Consider the toy corpus in Fig. 1. The English documents contain mentions of two people: Steven Paul Jobs and Mark Elliot Zuckerberg. Of course, the surface realization of Mr. Jobs' last name in English is also an ordinary nominal, hence the ambiguous mention string (absent context) in the second document. The Arabic document introduces an organization entity (Apple Inc.) along with proper and pronominal references to Mr. Jobs. Finally, the French document refers to Mr. Jobs by the honorific 'Monsieur,' and to

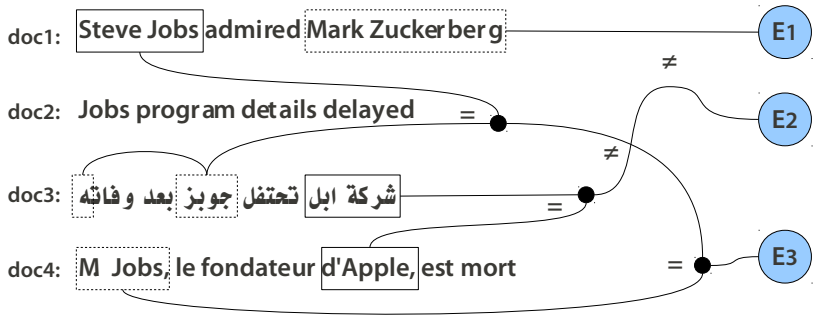


Figure 1: Clustering entity mentions across languages and documents. The toy corpus contains English (doc1 and doc2), Arabic (doc3), and French (doc4). Together, the documents make reference to three real-world entities, the identification of which is the primary objective of this work. We use a separately-trained system for within-document mention detection and coreference (indicated by the text boxes and intra-document links, respectively). Our experimental results are for Arabic-English only.

Apple without its corporate designation.

Our goal is to automatically produce the cross-lingual entity clusters $E_1$ (*Mark Elliot Zuckerberg*), $E_2$ (*Apple Inc.*), and $E_3$ (*Steven Paul Jobs*). Both the true number and characteristics of these entities are unobserved. Our models require two pre-processing steps: mention detection and within-document coreference/anaphora resolution, shown in Fig. 1 by the text boxes and intra-document links, respectively. For example, in doc3, a within-document coreference system would pre-link جوبز *joobz* 'Jobs' with the masculine pronoun ه *h* 'his'. In addition, the mention detector determines that the surface form "Jobs" in doc2 is not an entity reference. For this within-document pre-processing we use Serif (Ramshaw et al., 2011).[1]

Our models measure cross-lingual similarity of the coreference chains to make clustering decisions (• in Fig. 1). The similarity models (indicated by the $=$ and $\neq$ operators in Fig. 1) consider both mention string and context similarity (§2). We use the mention similarities as hard constraints, and the context similarities as soft constraints. In this work, we investigate two standard constrained clustering algorithms (§3). Our methods can be used to extend existing systems for mono-lingual entity clustering (also known as "cross-document coreference resolution") to the cross-lingual setting.

[1]Serif is a commercial system that assumes each document contains only one language. Currently, there are no publicly available within-document coreference systems for Arabic and many other languages. To remedy this problem, the CoNNL-2012 shared task aims to develop multilingual coreference systems.

## 2 Mention and Context Similarity

Our goal is to create cross-lingual sets of co-referent mentions to real-world entities (people, places, organizations, etc.). In this paper, we adopt the following notation. Let $M$ be a set of distinct text mentions in a collection of documents; $C$ is a partitioning of $M$ into document-level sets of co-referent mentions (called *coreference chains*); $E$ is a partitioning of $C$ into sets of co-referent chains (called *entities*). Let $i, j$ be non-negative integers less than or equal to $|M|$ and $a, b$ be non-negative integers less than or equal to $|C|$. Our experiments use a separate within-document coreference system to create $C$, which is fixed. We will learn $E$, which has size no greater than $|C|$ since the set of mono-lingual chains is the largest valid partitioning.

We define accessor functions to access properties of mentions and chains. For any mention $m_i$, define the following functions: $lang(m_i)$ is the language; $doc(m_i)$ is the document containing $m_i$; $type(m_i)$ is the semantic type, which is assigned by the within-document coreference system. We also extract a set of mention contexts $S$, which are the sentences containing each mention (i.e., $|S| = |M|$).

We learn the partition $E$ by considering mention and context similarity, which are measured with separate component models.

### 2.1 Mention Similarity

We use separate methods for within- and cross-language mention similarity. The pairwise similarity

| 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 | ي → ∅ | ء → ∅ | |
| **English Rules** | | | |
| **k** → c | **p** → b | **x** → ks | **e,i,o,u** → ∅ |

Table 1: English-Arabic mapping rules to a common orthographic representation. "∅" indicates a null mapping. For English, we also lowercase and remove determiners and punctuation. For Arabic, we remove the determiner ال *Al* 'the' and the elongation character *tatwil* 'ـ'.

of any two mentions $m_i$ and $m_j$ is:

$$sim(m_i, m_j) = \begin{cases} jaro\text{-}winkler(m_i, m_j) & \text{if } lang(m_i) = lang(m_j) \\ maxent(m_i, m_j) & \text{otherwise} \end{cases}$$

**Jaro-Winkler Distance (within-language)** If $lang(m_i) = lang(m_j)$, we use the Jaro-Winkler edit distance (Porter and Winkler, 1997). Jaro-Winkler rewards matching prefixes, the empirical justification being that less variation typically occurs at the beginning of names.[2] The metric produces a score in the range [0,1], where 0 indicates equality.

**Maxent model (cross-language)** When $lang(m_i) \neq lang(m_j)$, then the two mentions might be in different writing systems. Edit distance calculations no longer apply directly. One solution would be full-blown transliteration (Knight and Graehl, 1998), followed by application of Jaro-Winkler. However, transliteration systems are complex and require significant training resources. We find that a simpler, low-resource approach works well in practice.

First, we deterministically map both languages to a common phonetic representation (Tbl. 1).[3] Next, we align the mention pairs with the Hungarian algorithm, which produces a word-to-word alignment $A_{m_i,m_j}$.[4] Finally, we build a simple binary Maxent classifier $p(y|m_i, m_j; \lambda)$ that extracts features from the aligned mentions (Tbl. 2). We learn the parameters $\lambda$ using a quasi-Newton procedure with $L_1$ (lasso) regularization (Andrew and Gao, 2007).

| | |
|---|---|
| OVERLAP | Active for each bigram in $cbigrams(m_{i,u}) \bigcup cbigrams(m_{j,v})$ |
| BIGRAM-DIFF-$m_i$ | Active for each bigram in $cbigrams(m_i) - cbigrams(m_j)$ |
| BIGRAM-DIFF-$m_j$ | Active for each bigram in $cbigrams(m_j) - cbigrams(m_i)$ |
| BIGRAM-LEN-DIFF | Value of abs$(size(cbigrams(m_i) - cbigrams(m_j)))$ |
| BIG-EDIT-DIST | Count of token pairs with $Lev(m_{i,u}, m_{j,v}) > 3.0$ |
| TOTAL-EDIT-DIST | Sum of aligned token edit distances |
| LENGTH | Active for one of: $len(m_i) > len(m_j)$ or $len(m_i) < len(m_j)$ or $len(m_i) = len(m_j)$ |
| LENGTH-DIFF | abs$(len(m_i) - len(m_j))$ |
| SINGLETON | Active if $len(m_i) = 1$ |
| SINGLETON-PAIR | Active if $len(m_i) = len(m_j) = 1$ |

Table 2: Cross-language Maxent feature templates for a *whitespace-tokenized* mention pair $\langle m_i, m_j \rangle$ with alignment $A_{m_i,m_j}$. Let $(u, v) \in A_{m_i,m_j}$ indicate aligned token indices. Define the following functions for strings: $cbigrams(\cdot)$ returns the set of character bigrams; $len(\cdot)$ is the token length; $Lev(\cdot, \cdot)$ is the Levenshtein edit distance between two strings. Prior to feature extraction, we add unique start and end symbols to the mention strings.

## 2.2 Context Mapping and Similarity

Mention strings alone are not always sufficient for disambiguation. Consider again the simple example in Fig. 1. Both doc3 and doc4 reference "Steve Jobs" and "Apple" in the same contexts. Context co-occurence and/or similarity can thus disambiguate these two entities from other entities with similar references (e.g., "Steve Jones" or "Apple Corps"). As with the mention strings, the contexts may originate in different writing systems. We consider both high- and low-resource approaches for mapping contexts to a common representation.

[2] For multi-token names, we sort the tokens prior to computing the score, as suggested by Christen (2006).

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

[4] The Hungarian algorithm finds an optimal minimum-cost alignment. For pairwise costs between tokens, we used the Levenshtein edit distance

**Machine Translation (MT)** For the high-resource setting, if $lang(m_i) \neq$ English, then we translate both $m_i$ and its context $s_i$ to English with an MT system. We use Phrasal (Cer et al., 2010), a phrase-based system which, like most public MT systems, lacks a transliteration module. We believe that this approach yields the most accurate context mapping for high-resource language pairs (like English-Arabic).

**Polylingual Topic Model (PLTM)** The polylingual topic model (PLTM) (Mimno et al., 2009) is a generative process in which document tuples—groups of topically-similar documents—share a topic distribution. The tuples need not be sentence-aligned, so training data is easier to obtain. For example, one document tuple might be the set of Wikipedia articles (in all languages) for Steve Jobs.

Let $D$ be a set of document tuples, where there is one document in each tuple for each of $L$ languages. Each language has vocabulary $V_l$ and each document $d_t^l$ has $N_t^l$ tokens. We specify a fixed-size set of topics $K$. The PLTM generates the document tuples as follows:

| Polylingual Topic Model | |
|---|---|
| $\theta_t \sim \text{Dir}(\alpha^K)$ | [cross-lingual tuple-topic prior] |
| $\phi_k^l \sim \text{Dir}(\beta^{V_l})$ | [word-topic prior] |
| for each token $w_{t,n}^l$ with $n = \{1, \ldots, N_t^l\}$: | |
| $z_{t,n} \sim \text{Mult}(\theta_t)$ | |
| $w_{t,n}^l \sim \text{Mult}(\phi_{z_{t,n}}^l)$ | |

For cross-lingual context mapping, we infer the 1-best topic assignments for each token in all $S$ mention contexts. This technique reduces $V_l = k$ for all $l$. Moreover, all languages have a common vocabulary: the set of $K$ topic indices. Since the PLTM is not a contribution of this paper, we refer the interested reader to (Mimno et al., 2009) for more details.

After mapping each mention context to a common representation, we measure context similarity based on the choice of clustering algorithm.

## 3 Clustering Algorithms

We incorporate the mention and context similarity measures into a clustering framework. We consider two algorithms. The first is hierarchical agglomerative clustering (HAC), with which we assume basic familiarity (Manning et al., 2008). A shortcoming of HAC is that a stop threshold must be tuned. To avoid this requirement, we also consider non-parametric probabilistic clustering in the form of a Dirichlet process mixture model (DPMM) (Antoniak, 1974) .

Both clustering algorithms can be modified to accommodate pairwise constraints. We have observed better results by encoding mention similarity as a hard constraint. Context similarity is thus the cluster distance measure.[5]

To turn the Jaro-Winkler distance into a hard boolean constraint, we tuned a threshold $\eta$ on held-out data, i.e., $jaro\text{-}winkler(m_i, m_j) \leq \eta \Rightarrow m_i = m_j$. Likewise, the Maxent model is a binary classifier, so $p(y = 1|m_i, m_j; \lambda) > 0.5 \Rightarrow m_i = m_j$.

In both clustering algorithms, any two chains $C_a$ and $C_b$ cannot share the same cluster assignment if:

1. **Document origin**: $doc(C_a) = doc(C_b)$
2. **Semantic type**: $type(C_a) \neq type(C_b)$
3. **Mention Match**: $sim(m_i, m_j) = false$, where $m_i = repr(C_a)$ and $m_j = repr(C_b)$.

The deterministic accessor function $repr(C_a)$ returns the *representative mention* of a chain. The heuristic we used was "first mention": the function returns the earliest mention that appears in the associated document. In many languages, the first mention is typically more complete than later mentions. This heuristic also makes our system less sensitive to within-document coreference errors.[6] The representative mention only has special status for mention similarity: context similarity considers *all* mention contexts.

### 3.1 Constrained Hierarchical Clustering

HAC iteratively merges the "nearest" clusters according to context similarity. In our system, each cluster context is a bag of words $W$ formed from the contexts of all coreference chains in that cluster. For each word in $W$ we estimate a unigram Entity Language Model (ELM) (Raghavan et al., 2004):

$$P(w) = \frac{count_W(w) + \rho P_V(w)}{\sum_{w'} count_W(w') + \rho}$$

$P_V(w)$ is the unigram probability in *all* contexts in the corpus[7] and $\rho$ is a smoothing parameter. For any

[5] Specification of a combined similarity measure is an interesting direction for future work.

[6] These constraints are similar to the *pair-filters* of Mayfield et al. (2009).

[7] Recall that after context mapping, all languages have a common vocabulary $V$.

two entity clusters $E_a$ and $E_b$, the distance between $P_{E_a}$ and $P_{E_b}$ is given by a metric based on the Jensen-Shannon Divergence (JSD) (Endres and Schindelin, 2003):

$$dist(P_{E_a}, P_{E_b}) = \sqrt{2 \cdot JSD(P_{E_a}||P_{E_b})}$$
$$= \sqrt{KL(P_{E_a}||M) + KL(M||P_{E_b})}$$

where $KL(P_{E_a}||M)$ is the Kullback-Leibler divergence and $M = \frac{1}{2}(P_{E_a} + P_{E_b})$.

We initialize HAC to $E = C$, i.e., the initial clustering solution is just the set of all coreference chains. Then we remove all links in the HAC proximity matrix that violate pairwise cannot-link constraints. During clustering, we do not merge $E_a$ and $E_b$ if *any* pair of chains violates a cannot-link constraint. This procedure propagates the cannot-link constraints (Klein et al., 2002). To output $E$, we stop clustering when the minimum JSD exceeds a stop threshold $\gamma$, which is tuned on a development set.

### 3.2 Constrained Dirichlet Process Mixture Model (DPMM)

Instead of tuning a parameter like $\gamma$, it would be preferable to let the data dictate the number of entity clusters. We thus consider a non-parametric Bayesian mixture model where the mixtures are multinomial distributions over the entity contexts $S$. Specifically, we consider a DPMM, which automatically infers the number of mixtures. Each $C_a$ has an associated mixture $\theta_a$:

$$C_a|\theta_a \sim \text{Mult}(\theta_a)$$
$$\theta_a|G \sim \text{G}$$
$$G|\alpha, G_0 \sim \text{DP}(\alpha, G_0)$$
$$\alpha \sim \text{Gamma}(1, 1)$$

where $\alpha$ is the concentration parameter of the DP prior and $G_0$ is the base distribution with support $V$. For our experiments, we set $G_0 = \text{Dir}(\pi_1, \ldots, \pi_V)$, where $\pi_i = P_V(w_i)$.

For inference, we use the Gibbs sampler of Vlachos et al. (2009), which can incorporate pairwise constraints. The sampler is identical to a standard collapsed, token-based sampler, except the conditional probability $p(E_a = E|E_{-a}, C_a) = 0$ if $C_a$ cannot be merged with the chains in cluster $E$. This property makes the model non-exchangeable, but in practice non-exchangeable models are sometimes useful (Blei and Frazier, 2010). During sampling, we also learn $\alpha$ using the auxiliary variable procedure of West (1995), so the only fixed parameters are those of the vague Gamma prior. However, we found that these hyperparameters were not sensitive.

## 4 Training Data and Procedures

We trained our system for Arabic-English cross-lingual entity clustering.[8]

**Maxent Mention Similarity** The Maxent mention similarity model requires a parallel name list for training. Name pair lists can be obtained from the LDC (e.g., LDC2005T34 contains nearly 450,000 parallel Chinese-English names) or Wikipedia (Irvine et al., 2010). We extracted 12,860 name pairs from the parallel Arabic-English translation treebanks,[9] although our experiments show that the model achieves high accuracy with significantly fewer training examples. We generated a uniform distribution of training examples by running a Bernoulli trial for each aligned name pair in the corpus. If the coin was heads, we replaced the English name with another English name chosen randomly from the corpus.

**MT Context Mapping** For the MT context mapping method, we trained Phrasal with all data permitted under the NIST OpenMT Ar-En 2009 constrained track evaluation. 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. In addition to the baseline Phrasal feature set, we used the lexicalized re-ordering model of Galley and Manning (2008).

**PLTM Context Mapping** For PLTM training, we formed a corpus of 19,139 English-Arabic topically-aligned Wikipedia articles. 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). To increase vocabulary coverage for our ACE2008 evaluation corpus, we added 20,000 document singletons from the ACE2008 training corpus. The

[8] We tokenized all English documents with packages from the Stanford parser (Klein and Manning, 2003). For Arabic documents, we used Mada (Habash and Rambow, 2005) for orthographic normalization and clitic segmentation.

[9] LDC Catalog numbers LDC2009E82 and LDC2009E88.

topically-aligned tuples served as "glue" to share topics between languages, while the ACE documents distribute those topics over in-domain vocabulary.[10]

We used the PLTM implementation in Mallet (McCallum, 2002). We ran the sampler for 10,000 iterations and set the number of topics $K = 512$.

## 5 Task Evaluation Framework

Our experimental design is a cross-lingual extension of the standard cross-document coreference resolution task, which appeared in ACE2008 (Strassel et al., 2008; NIST, 2008). We evaluate name (NAM) mentions for cross-lingual person (PER) and organization (ORG) entities. Neither the number nor the attributes of the entities are known (i.e., the task does not include a knowledge base). We report results for both gold and automatic within-document mention detection and coreference resolution.

**Evaluation Metrics** We use *entity-level* evaluation metrics, i.e., we evaluate the $E$ entity clusters rather than the mentions. For the gold setting, we report:

- $B^3$ (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.
- NVI (Reichart and Rappoport, 2009): Information-theoretic measure that utilizes the entropy of the clusters and their mutual information. Unlike the commonly-used Variation of Information (VI) metric, normalized VI (NVI) is not sensitive to the size of the data set.

For the automatic setting, we must apply a different metric since the number of system chains may differ from the reference. We use $B^3_{\text{sys}}$ (Cai and Strube, 2010), a variant of $B^3$ that was shown to penalize both twinless reference chains and spurious system chains more fairly.

**Evaluation Corpus** The automatic evaluation of cross-lingual coreference systems requires annotated multilingual corpora. 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 seven genres sampled from independent sources over the course of a decade (Tbl. 3). The corpus provides gold mono-lingual cross-document coreference annotations for both PER and ORG entities. Using these annotations as a starting point, we found and annotated 216 cross-lingual entities.[11]

| | Docs | Tokens | Entities | Chains | Mentions |
|---|---|---|---|---|---|
| Arabic | 412 | 178,269 | 2,594 | 4,216 | 9,222 |
| English | 414 | 246,309 | 2,278 | 3,950 | 9,140 |

Table 3: ACE2008 evaluation corpus PER and ORG entity statistics. Singleton chains account for 51.4% of the Arabic data and 46.2% of the English data. Just 216 entities appear in both languages.

Because a similar corpus did not exist for development, we split the evaluation corpus into development and test sections. However, the usual method of splitting by document would not confine all mentions of each entity to one side of the split. We thus split the corpus by *global entity id*. We assigned one-third of the entities to development, and the remaining two-thirds to test.

## 6 Comparison to Related Tasks and Work

Our modeling techniques and task formulation can be viewed as cross-lingual extensions to cross-document coreference resolution. The classic work on this task was by Bagga and Baldwin (1998b), who adapted the Vector Space Model (VSM) (Salton et al., 1975). Gooi and Allan (2004) found effective algorithmic extensions like agglomerative clustering. Successful feature extensions to the VSM for cross-document coreference 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. Fleischman and Hovy (2004) added a discriminative pairwise mention classifier to a VSM-like model, much as we do. More

[10] 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.

[11] The annotators were the first author and another fluent speaker of Arabic. The annotations, corrections, and corpus split are available at http://www.spencegreen.com/research/.

recent work has considered new models for web-scale corpora (Rao et al., 2010; Singh et al., 2011).

Cross-document work on languages other than English is 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 within-document coreference system of Florian et al. (2004). They clustered the entities incrementally using a binary 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 name uniqueness. Finally, Sayeed et al. (2009) translated Arabic name mentions to English and then formed clusters greedily using pairwise matching.

To our knowledge, the cross-lingual entity clustering task is novel. However, there is significant prior work on similar tasks:

- **Multilingual coreference resolution**: Adapt English within-document coreference models to other languages (Harabagiu and Maiorano, 2000; Florian et al., 2004; Luo and Zitouni, 2005).
- **Named entity translation**: For a non-English document, produce an inventory of entities in English. An ACE2007 pilot task (Song and Strassel, 2008).
- **Named entity clustering**: Assign semantic types to text mentions (Collins and Singer, 1999; Elsner et al., 2009).
- **Cross-language name search / entity linking**: Match a *single* query name against a list of known multilingual names (knowledge base). A track in the 2011 NIST Text Analysis Conference (TAC-KBP) evaluation (Aktolga et al., 2008; McCarley, 2009; Udupa and Khapra, 2010; McNamee et al., 2011).

Our work incorporates elements of the first three tasks. Most importantly, we avoid the key element of entity linking: a knowledge base.

## 7 Experiments

We performed intrinsic evaluations for both mention and context similarity. For context similarity, we analyzed mono-lingual entity clustering, which also facilitated comparison to prior work on the ACE2008 evaluation set. Our main results are for the new task: cross-lingual entity clustering.

| Genre | #Train | #Test | Accuracy(%) |
|---|---|---|---|
| wb | 125 | 16 | 87.5 |
| bn | 2,720 | 340 | 95.6 |
| nw | 7,443 | 930 | 96.6 |
| all | 10,288 | 1,286 | 97.1 (+7.55) |

Table 4: Cross-lingual mention matching accuracy [%]. The training data contains names from three genres: broadcast news (bn), newswire (nw), and weblog (wb). We used the full training corpus (all) for the cross-lingual clustering experiments, but the model achieved high accuracy with significantly fewer training examples (e.g., bn).

| | CEAF↑ | NVI↓ | $B^3$ ↑ | | | |
|---|---|---|---|---|---|---|
| | | | #hyp | P | R | F1 |
| **Mono-lingual Arabic** (#gold=1,721) | | | | | | |
| HAC | 87.2 | 0.052 | 1,669 | 89.8 | 89.8 | 89.8 |
| **Mono-lingual English** (#gold=1,529) | | | | | | |
| HAC | 88.5 | 0.042 | 1,536 | 93.7 | 89.0 | 91.4 |

Table 5: Mono-lingual entity clustering evaluation (test set, *gold* within-document processing). Higher scores (↑) are better for CEAF and $B^3$, whereas lower (↓) is better for NVI. #gold indicates the number of reference entities, whereas #hyp is the size of $E$.

### 7.1 Intrinsic Evaluations

**Cross-lingual Mention Matching** We created a random 80/10/10 (train, development, test) split of the Maxent training corpus and evaluated binary classification accuracy (Tbl. 4). Of the mis-classified examples, we observed three major error types. First, the model learns that high edit distance is predictive of a mismatch. 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 كيزرسلوترن (phonetic mapping: *af s kazrslawtrn*) versus "FC Kaiserslautern." Finally, names that require translation are problematic. For example, the classifier produces a false negative for ⟨God, *gd*⟩ $\stackrel{?}{=}$ ⟨ألله, *allh*⟩.

| #gold = 3,057 | CEAF↑ | NVI↓ | $B^3$ ↑ | | | | $B^3_{\text{target}}$ ↑ (#gold = 146) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | #hyp | P | R | F1 | #hyp | P | R | F1 |
| SINGLETON | 64.9 | 0.165 | 5,453 | **100.0** | 56.1 | 71.8 | 1,587 | **100.0** | 9.20 | 16.9 |
| NO-CONTEXT | 57.4 | 0.136 | 2,216 | 65.6 | 75.2 | 70.1 | 517 | 78.3 | 41.8 | 54.5 |
| HAC+MT | **79.8** | **0.070** | 2,783 | 84.4 | **86.4** | **85.4** | 310 | 91.7 | **69.1** | **78.8** |
| DPMM+MT | 74.3 | 0.122 | 3,649 | 89.3 | 64.1 | 74.6 | 634 | 93.3 | 24.3 | 38.6 |
| HAC+PLTM | 72.1 | 0.110 | 2,746 | 76.9 | 77.6 | 77.3 | 506 | 84.4 | 44.6 | 58.4 |
| DPMM+PLTM | 57.2 | 0.180 | 2,609 | 64.0 | 62.8 | 63.4 | 715 | 73.9 | 22.2 | 34.1 |

Table 6: Cross-lingual entity clustering (test set, *gold* within-document processing). $B^3_{\text{target}}$ is the standard $B^3$ metric applied to the subset of target cross-lingual entities in the test set. For CEAF and $B^3$, SINGLETON is the stronger baseline due to the high proportion of singleton entities in the corpus. Of course, cross-lingual entities have at least two chains, so NO-CONTEXT is a better baseline for cross-lingual clustering.

**Mono-lingual Entity Clustering** For comparison, we also evaluated our system on a standard mono-lingual cross-document coreference task (Arabic and English) (Tbl. 5). We configured the system with HAC clustering and Jaro-Winkler (within-language) mention similarity. We built mono-lingual ELMs for context similarity.

We used two baselines:

- SINGLETON: $E = C$, i.e., the cross-lingual clustering solution is just the set of mono-lingual coreference chains. This is a common baseline for mono-lingual entity clustering (Baron and Freedman, 2008).
- NO-CONTEXT: We run HAC with $\rho = \infty$. Therefore, $E$ is the set of fully-connected components in $C$ subject to the pairwise constraints.

For HAC, we manually tuned the stop threshold $\gamma$, the Jaro-Winkler threshold $\eta$, and the ELM smoothing parameter $\rho$ on the development set. For the DPMM, no development tuning was necessary, and we evaluated a single sample of $E$ taken after 3,000 iterations.

To our knowledge, Baron and Freedman (2008) reported the only previous results on the ACE2008 data set. However, they only gave gold results for English, and clustered the entire evaluation corpus (test+development). To control for the effect of within-document errors, we considered their gold input (mention detection and within-document coreference resolution) results. They reported $B^3$ for the two entity types separately: ORG (91.5% F1) and PER (94.3% F1). The different experimental designs preclude a precise comparison, but the accuracy of the two systems are at least in the same range.

| #gold = 3,057 | $B^3_{sys}$ ↑ | | | |
|---|---|---|---|---|
| | #hyp | P | R | F1 |
| SINGLETON | 7,655 | **100.0** | 57.1 | 72.7 |
| NO-CONTEXT | 2,918 | 63.3 | 71.1 | 67.0 |
| HAC+MT | 3,804 | 75.6 | **77.8** | **76.7** |
| DPMM+MT | 4,491 | 77.1 | 62.5 | 69.0 |
| HAC+PLTM | 6,353 | 94.1 | 62.8 | 75.3 |
| DPMM+PLTM | 3,522 | 64.6 | 62.0 | 63.3 |

Table 7: Cross-lingual entity clustering (test set, *automatic* (Serif) within-document processing). For HAC, we used the same parameters as the gold setting.

### 7.2 Cross-lingual Entity Clustering

We evaluated four system configurations on the new task: HAC+MT, HAC+PLTM, DPMM+MT, and DPMM+PLTM. First, we established an upper bound by assuming gold within-document mention detection and coreference resolution (Tbl. 6). This setting isolated the new cross-lingual clustering methods from within-document processing errors. Then we evaluated with Serif (automatic) within-document processing (Tbl. 7). This second experiment replicated an application setting. We used the same baselines and tuning procedures as in the mono-lingual clustering experiment.

**Results** In the gold setting, HAC+MT produces the best results, as expected. The dimensionality reduction of the vocabulary imposed by PLTM significantly reduces accuracy, but HAC+PLTM still exceeds the

baseline. We tried increasing the number of PLTM topics $k$, but did not observe an improvement in task accuracy. For both context-mapping methods, the DPMM suffers from low-recall. Upon inspection, the clustering solution of DPMM+MT contains a high proportion of singleton hypotheses, suggesting that the model finds lower similarity in the presence of a larger vocabulary. When the context vocabulary consists of PLTM topics, larger clusters are discovered (DPMM+PLTM).

The effect of dimensionality reduction is also apparent in the clustering solutions of the PLTM models. For example, for the Serif output, DPMM+PLTM produces a cluster consisting of "White House", "Senate", "House of Representatives", and "Parliament". Arabic mentions of the latter three entities pass the pairwise mention similarity constraints due to the word مجلس 'council', which appears in text mentions for all three legislative bodies. A cross-language matching error resulted in the linking of "White House", and the reduced granularity of the contexts precluded further disambiguation. Of course, these entities probably appear in similar contexts.

The caveat with the Serif results in Tbl. 7 is that 3,251 of the 7,655 automatic coreference chains are *not* in the reference. Consequently, the evaluation is dominated by the penalty for spurious system coreference chains. Nonetheless, all models except for DPMM+PLTM exceed the baselines, and the relationships between models depicted in the gold experiments hold for the this setting.

## 8 Conclusion

Cross-lingual entity clustering is a natural step toward more robust natural language understanding. We proposed pipeline models that make clustering decisions based on cross-lingual similarity. We investigated two methods for mapping documents in different languages to a common representation: MT and the PLTM. Although MT may achieve more accurate results for some language pairs, the PLTM training resources (e.g., Wikipedia) are readily available for many languages. As for the clustering algorithms, HAC appears to perform better than the DPMM on our dataset, but this may be due to the small corpus size. The instance-level constraints represent tendencies that could be learned from larger amounts of data. With more data, we might be able to relax the constraints and use an exchangeable DPMM, which might be more effective. Finally, we have shown that significant quantities of within-document errors cascade into the cross-lingual clustering phase. As a result, we plan a model that clusters the mentions directly, thus removing the dependence on within-document coreference resolution.

In this paper, we have set baselines and proposed models that significantly exceeded those baselines. The best model improved upon the cross-lingual entity baseline by 24.3% F1. This result was achieved without a knowledge base, which is required by previous approaches to cross-lingual entity linking. More importantly, our techniques can be used to extend existing cross-document entity clustering systems for the increasingly multilingual web.

**Acknowledgments** We thank Jason Eisner, David Mimno, Scott Miller, Jim Mayfield, and Paul McNamee for helpful discussions. This work was started during the SCALE 2010 summer workshop at Johns Hopkins. The first author is supported by a National Science Foundation Graduate Fellowship.

## 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*.

C. E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

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*.

D. Blei and P. Frazier. 2010. Distance dependent Chinese restaurant processes. In *ICML*.

J. Cai and M. Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*.

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*.

E. Charniak. 2001. Unsupervised learning of name structure from coreference data. In *NAACL*.

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.

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *EMNLP*.

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

M. Elsner, E. Charniak, and M. Johnson. 2009. Structured generative models for unsupervised named-entity clustering. In *HLT-NAACL*.

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

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*.

C. H. Gooi and J. Allan. 2004. Cross-document coreference on a large scale corpus. In *HLT-NAACL*.

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

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

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

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *ACL*.

D. Klein, S. D. Kamvar, and C. D. Manning. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*.

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

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*.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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*.

A. K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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

P. McNamee, J. Mayfield, D. Lawrie, D.W. Oard, and D. Doermann. 2011. Cross-language entity linking. In *IJCNLP*.

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*.

L. Ramshaw, E. Boschee, M. Freedman, J. MacBride, R. Weischedel, and A. Zamanian. 2011. SERIF language processing—effective trainable language understanding. In J. Olive et al., editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 636–644. Springer.

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

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

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *CACM*, 18:613–620, November.

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*.

A. Vlachos, A. Korhonen, and Z. Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics*.

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.

M. West. 1995. Hyperparameter estimation in Dirichlet process mixture models. Technical report, Duke University.