Transferring Coreference Resolvers with Posterior Regularization

André F. T. Martins[*†]
[*]Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[†]Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
atm@priberam.pt

Abstract

We propose a cross-lingual framework for learning coreference resolvers for resource-poor target languages, given a resolver in a source language. Our method uses word-aligned bitext to project information from the source to the target. To handle task-specific costs, we propose a softmax-margin variant of posterior regularization, and we use it to achieve robustness to projection errors. We show empirically that this strategy outperforms competitive cross-lingual methods, such as delexicalized transfer with bilingual word embeddings, bitext direct projection, and vanilla posterior regularization.

1 Introduction

The goal of **coreference resolution** is to find the mentions in text that refer to the same discourse entity. While early work focused primarily on English (Soon et al., 2001; Ng and Cardie, 2002), efforts have been made toward multilingual systems, this being addressed in recent shared tasks (Recasens et al., 2010; Pradhan et al., 2012). However, the lack of annotated data hinders rapid system deployment for new languages. Unsupervised methods (Haghighi and Klein, 2007; Ng, 2008) and rule-based approaches (Raghunathan et al., 2010) avoid this data annotation bottleneck, but they often require complex generative models or expert linguistic knowledge.

We propose **cross-lingual coreference resolution** as a way of transferring information from a rich-resource language to build coreference resolvers for languages with scarcer resources; as a testbed, we transfer from English to Spanish and to Brazilian Portuguese. We build upon the recent successes of cross-lingual learning in NLP, which proved quite effective in several structured prediction tasks, such as POS tagging (Täckström et al., 2013), named entity recognition (Wang and Manning, 2014), dependency parsing (McDonald et al., 2011), semantic role labeling (Titov and Klementiev, 2012), and fine-grained opinion mining (Almeida et al., 2015). The potential of these techniques, however, has never been fully exploited in coreference resolution (despite some existing work, reviewed in §6, but none resulting in an end-to-end coreference resolver).

We bridge this gap by proposing a simple learning-based method with weak supervision, based on **posterior regularization** (Ganchev et al., 2010). We adapt this framework to handle softmax-margin objective functions (Gimpel and Smith, 2010), leading to **softmax-margin posterior regularization** (§4). This step, while fairly simple, opens the door for incorporating task-specific cost functions, which are important to manage the precision/recall trade-offs in coreference resolution systems. We show that the resulting problem involves optimizing the difference of two cost-augmented log-partition functions, making a bridge with supervised systems based on **latent coreference trees** (Fernandes et al., 2012; Durrett and Klein, 2013), reviewed in §3. Inspired by this idea, we consider a simple **penalized variant** of posterior regularization that tunes the Lagrange multipliers directly, bypassing the saddle-point problem of existing EM and alternating stochastic gradient algorithms (Ganchev et al., 2010; Liang et al., 2009). Experiments (§5) show that the proposed method outperforms commonly used cross-lingual approaches, such as delexicalized transfer with bilingual embeddings, direct projection, and "vanilla" posterior regularization.

2 Architecture and Experimental Setup

Our methodology, outlined as Algorithm 1, is inspired by the recent work of Ganchev and Das (2013) on cross-lingual learning of sequence models. For simplicity, we call the source and tar-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1427–1437, Beijing, China, July 26-31, 2015.

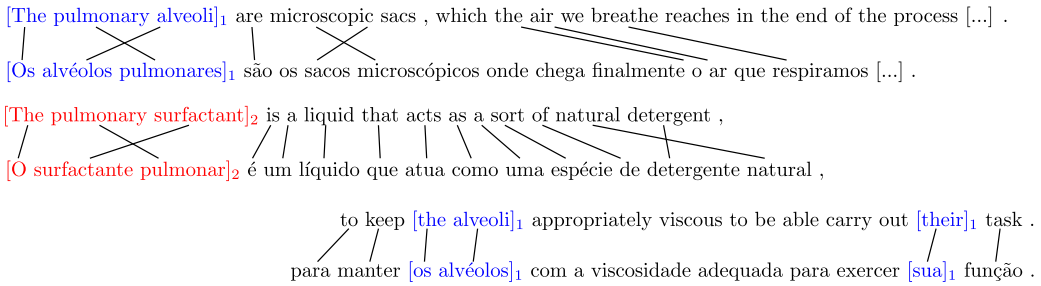

Figure 1: Excerpt of a bitext document with automatic coreference annotations (from FAPESP). The English side had its coreferences resolved by a state-of-the-art system (Durrett and Klein, 2013). The predicted coreference chains {*The pulmonary alveoli*, *the alveoli*, *their*} and {*The pulmonary surfactant*} are then projected to the Portuguese side, via word alignments.

Algorithm 1 Cross-Lingual Coreference Resolution via Softmax-Margin Posterior Regularization

Input: Source coreference system $\mathcal{S}^e$, parallel data $\mathcal{D}^e$ and $\mathcal{D}^f$, posterior constraints $\mathcal{Q}$.
Output: Target coreference system $\mathcal{S}^f$.
1: $\mathcal{D}^{e \leftrightarrow f} \leftarrow \textsc{RunWordAligner}(\mathcal{D}^e, \mathcal{D}^f)$
2: $\widehat{\mathcal{D}}^e \leftarrow \textsc{RunCoref}(\mathcal{S}^e, \mathcal{D}^e)$
3: $\widehat{\mathcal{D}}^f \leftarrow \textsc{ProjectAndFilterEntities}(\mathcal{D}^{e \leftrightarrow f}, \widehat{\mathcal{D}}^e)$
4: $\mathcal{S}^f \leftarrow \textsc{LearnCorefWithSoftMargPR}(\widehat{\mathcal{D}}^f, \mathcal{Q})$

get languages English (e) and "foreign" (f), respectively, and we assume the existence of parallel documents on the two languages (bitext).

The first two steps (lines 1–2) run a word aligner and label the source side of the parallel data with a pre-trained English coreference system. Afterwards, the predicted English entities are projected to the target side of the parallel data (line 3), inducing an automatic (and noisy) training dataset for the foreign language. Finally, a coreference system is trained in this dataset with the aid of softmax-margin posterior regularization (line 4).

We next detail all the datasets and tools involved in our experimental setup. Table 1 provides a summary, along with some statistics.

Parallel Data. As parallel data, we use a sentence-aligned trilingual (English-Portuguese-Spanish) parallel corpus based on the scientific news Brazilian magazine *Revista Pesquisa FAPESP*, collected by Aziz and Specia (2011).[1] We preprocessed this dataset as follows. We labeled the English side with the Berkeley Coreference Resolution system v1.0, using the provided English model (Durrett and Klein, 2013). Then, we computed word alignments using the Berkeley aligner (Liang et al., 2006), intersected them and filtered out all the alignments whose confidence is below 0.95. After this, we projected English mentions to the target side using the maximal span heuristic of Yarowsky et al. (2001). We filtered out documents where more than 15% of the mentions were not aligned. At this point, we obtained an automatically annotated corpus $\widehat{\mathcal{D}}^f$ in the target language. Figure 1 shows a small excerpt where all mentions were correctly projected. In practice, not all documents are so well behaved: in the English-Portuguese parallel data, only 200,175 out of the original 271,122 mentions (about 73.8%) were conserved after the projection step. In Spanish, this number drops to 69.9%.

[1]We found that other commonly used parallel data (such as Europarl or the UN corpus) have a predominance of direct speech that is not suitable for our newswire test domain, so we decided not to use these data.

Dataset	# Doc.	# Sent.	# Tok.
EN OntoNotes (train)	2,374	48,762	1,007,359
EN OntoNotes (dev)	303	6,894	136,257
EN OntoNotes (test)	322	8,262	152,728
ES FAPESP (aligned)	2,704	142,633	3,840,936
ES AnCora (train)	875	8,999	295,276
ES AnCora (dev)	140	1,417	46,167
ES AnCora (test)	168	1,704	53,042
PT FAPESP (aligned)	2,823	166,719	4,538,147
PT Summ-It (train)	30	469	11,771
PT Summ-It (dev)	7	111	2,983
PT Summ-It (test)	13	257	6,491

Table 1: Corpus statistics. EN, ES, and PT denote English, Spanish, and Portuguese, respectively.

Monolingual Data. We also use monolingual data for validation and comparison with supervised systems. The Berkeley Coreference Resolution system is trained in the English OntoNotes dataset used in the CoNLL 2011 shared task; this dataset is also used to train delexicalized models.

For Spanish, we use the AnCora dataset (Recasens and Martí, 2010) provided in the SemEval 2010 coreference task, which we preprocessed as follows. We split all MWEs into individual tokens (for consistency with the other corpora). We also removed the extra gap tokens associated with zero-anaphoric relations, and the anaphoric annotations associated with relative pronouns (*e.g.*, in "*[una central de ciclo combinado [que]*$_1$ *debe empezar*

a funcionar en mayo del 2002]$_1$" we removed the nested mention *[que]*$_1$), since these are not annotated in the English dataset.

For Portuguese, we used the Summ-It 3.0 corpus (Collovini et al., 2007), which contains 50 documents annotated with coreferences, from the science section of the *Folha de São Paulo* newspaper. This dataset is much smaller than OntoNotes and AnCora, as shown in Table 1. We split the data into train, development, and test partitions.

For both Spanish and Portuguese, we obtained automatic POS tags and dependency parses by using TurboParser (Martins et al., 2013).

3 Coreference Resolution

3.1 Problem Definition and Prior Work

In coreference resolution, we are given a set of **mentions** $\mathcal{M} := \{m_1, \ldots, m_M\}$, and the goal is to cluster them into discourse **entities**, $\mathcal{E} := \{e_1, \ldots, e_E\}$, where each $e_j \subseteq \mathcal{M}$ and $e_j \neq \varnothing$. The set $\mathcal{E}$ must form a partition of $\mathcal{M}$, *i.e.*, we must have $\bigcup_{j=1}^{E} e_j = \mathcal{M}$, and $e_i \cap e_j = \varnothing$ for $i \neq j$.

A variety of approaches have been proposed to this problem, including entity-centric models (Haghighi and Klein, 2010; Rahman and Ng, 2011; Durrett et al., 2013), pairwise models (Bengtson and Roth, 2008; Versley et al., 2008), greedy rule-based methods (Raghunathan et al., 2010), and mention-ranking decoders (Denis and Baldridge, 2008; Durrett and Klein, 2013). We chose to base our coreference resolvers on this last class of methods, which permit efficient decoding by shifting from entity clusters to **latent coreference trees**. In particular, the inclusion of lexicalized features by Durrett and Klein (2013) yields nearly state-of-the-art performance with surface information only. Given that our goal is to prototype resolvers for resource-poor languages, this model is a good fit—we next describe it in detail.

3.2 Latent Coreference Tree Models

Let x be a document containing M mentions, sorted from left to right. We associate to the mth mention a random variable $y_m \in \{0, 1, \ldots, m-1\}$ to denote its **antecedent**, where the value $y_m = 0$ means that m is a singleton or starts a new coreference chain. We denote by $\mathcal{Y}(x)$ the set of coreference trees that can be formed by linking mentions to their antecedents; we represent each tree as a vector $y := \langle y_1, \ldots, y_M \rangle$. Note that each tree y induces a unique clustering $\mathcal{E}$, but that this map is many-to-one, *i.e.*, different trees may correspond to the same set of entity clusters. We denote by $\mathcal{Y}(\mathcal{E})$ the set of trees that are consistent with a given clustering $\mathcal{E}$.

We model the probability distribution $p(y|x)$ as an arc-factored log-linear model:

$$p_{\boldsymbol{w}}(y|x) \propto \exp\left(\sum_{m=1}^{M} \boldsymbol{w}^\top \boldsymbol{f}(x, m, y_m)\right), \quad (1)$$

where $\boldsymbol{w}$ is a weight vector, and each $\boldsymbol{f}(x, m, y_m)$ is a local feature vector that depends on the document x, the mention m, and its candidate antecedent y_m. This model permits a cheap computation of the most likely tree $\widehat{y} := \arg\max_{y \in \mathcal{Y}(x)} p_{\boldsymbol{w}}(y|x)$: simply compute the best antecedent independently for each mention, and collect them to form a tree. An analogous procedure can be employed to compute the posterior marginals $p_{\boldsymbol{w}}(y_m|x)$ for every mention m.

Gold coreference tree annotations are rarely available; datasets usually consist of documents annotated with entity clusters, $\{\langle x^{(n)}, \mathcal{E}^{(n)} \rangle\}_{n=1}^{N}$. Durrett and Klein (2013) proposed to learn the probabilistic model in Eq. 1 by maximizing conditional log-likelihood, treating the coreference trees as latent variables. They also found advantageous to incorporate a **cost function** $\ell(y, \mathcal{Y}(\mathcal{E}))$, measuring the extent to which a prediction y differs from the ones that are consistent with the gold entity set $\mathcal{E}$.[2] Putting these pieces together, we arrive at the following loss function to be minimized:

$$L(\boldsymbol{w}) = -\sum_{n=1}^{N} \log\left(\sum_{y \in \mathcal{Y}(\mathcal{E}^{(n)})} p'_{\boldsymbol{w}}(y|x^{(n)})\right), \quad (2)$$

where $p'_{\boldsymbol{w}}$ is the **cost-augmented distribution**:

$$p'_{\boldsymbol{w}}(y|x) \propto p_{\boldsymbol{w}}(y|x) e^{\ell(y, \mathcal{Y}(\mathcal{E}))}. \quad (3)$$

The loss function in Eq. 2 can be seen as a probabilistic analogous of the hinge loss of support vector machines, and a model trained this way is called a **softmax-margin** CRF (Gimpel and Smith, 2010). Note that $L(\boldsymbol{w})$ is non-convex, corresponding to the difference of two log-partition functions (both convex on $\boldsymbol{w}$),

$$L(\boldsymbol{w}) = \sum_{n=1}^{N} \left(\log Z'(\boldsymbol{w}, x^{(n)}) - \log \widehat{Z}(\boldsymbol{w}, x^{(n)})\right); \quad (4)$$

above we denoted

$$Z'(\boldsymbol{w}, x) = \sum_{y \in \mathcal{Y}(x)} e^{\boldsymbol{w}^\top \boldsymbol{f}(x,y) + \ell(y, \mathcal{Y}(\mathcal{E}))} \quad (5)$$

$$\widehat{Z}(\boldsymbol{w}, x) = \sum_{y \in \mathcal{Y}(\mathcal{E})} e^{\boldsymbol{w}^\top \boldsymbol{f}(x,y)}, \quad (6)$$

[2] A precise definition of this cost is provided in §4.3.

where $\boldsymbol{f}(x, y) := \sum_{m=1}^{M} \boldsymbol{f}(x, m, y_m)$.[3] Evaluating the gradient of the loss in Eq. 4 requires computing marginals for the candidate antecedents of each mention, which can be done in a mention-synchronous fashion. This enables a simple stochastic gradient descent algorithm, which was the procedure taken by Durrett and Klein (2013).

Another way of regarding this framework, expressed through the marginalization in Eq. 2, is to "pretend" that the outputs we care about are the actual coreference trees, but that the datasets are only "weakly labeled" with the entity clusters. We build on this point of view in §4.1.

4 Cross-Lingual Coreference Resolution

We now adapt the framework above to learn coreference resolvers in a cross-lingual manner.

4.1 Softmax-Margin Posterior Regularization

In the weakly supervised case, the training data may only be partially labeled or contain annotation errors. For taking advantage of these data, we need a procedure that handles uncertainty about the missing data, and is robust to mislabelings. We describe next an approach based on posterior regularization (PR) that fulfills these requirements.

For ease of explanation, we introduce corpus-level counterparts for the variables in §3.2. We use bold capital letters $\mathbf{X} := \{x^{(1)}, \ldots, x^{(N)}\}$ and $\mathbf{Y} := \{y^{(1)}, \ldots, y^{(N)}\}$ to denote the documents and candidate coreference trees in our corpus. We denote by $p_{\boldsymbol{w}}(\mathbf{Y}|\mathbf{X}) := \prod_{n=1}^{N} p_{\boldsymbol{w}}(y|x^{(n)})$ the conditional distribution of trees over the corpus, induced by a model $\boldsymbol{w}$, and similarly for the cost-augmented distribution $p'_{\boldsymbol{w}}(\mathbf{Y}|\mathbf{X})$.

In PR, we define a vector $\boldsymbol{g}(\mathbf{X}, \mathbf{Y})$ of **corpus-level constraint features**, and a vector $\boldsymbol{b}$ of upper bounds for those features. We consider the family of distributions over $\mathbf{Y}$ (call it $\mathcal{Q}$) that satisfy these constraints in *a posteriori* expectation,

$$\mathcal{Q} := \{q \mid \mathbf{E}_q[\boldsymbol{g}(\mathbf{X}, \mathbf{Y})] \leq \boldsymbol{b}\}. \tag{7}$$

To make the analysis simpler, we assume that $\mathbf{0} \leq \boldsymbol{b} \leq \mathbf{1}$, and that for every j, $\min_{\mathbf{Y}} g_j(\mathbf{X}, \mathbf{Y}) = 0$ and $\max_{\mathbf{Y}} g_j(\mathbf{X}, \mathbf{Y}) = 1$, where the min/max above are over all possible coreference trees $\mathbf{Y}$ that can be build from the documents $\mathbf{X}$ in the corpus.[4] Under this assumption, the two extreme values of the upper bounds have a precise meaning: if $b_j = 0$, the jth feature becomes a hard constraint, (*i.e.*, any feasible distribution in $\mathcal{Q}$ will vanish outside $\{\mathbf{Y} \mid g_j(\mathbf{X}, \mathbf{Y}) = 0\}$), while $b_j = 1$ turns it into a vacuous feature.

We also make the usual assumption that the constraint features decompose over documents, $\boldsymbol{g}(\mathbf{X}, \mathbf{Y}) := \sum_{n=1}^{N} \boldsymbol{g}(x^{(n)}, y^{(n)})$; if this were not the case, decoding would be much harder, as the documents would be coupled.

In vanilla PR (Ganchev et al., 2010), one seeks the model $\boldsymbol{w}$ minimizing the Kullback-Leibler divergence between the set $\mathcal{Q}$ and the distribution $p_{\boldsymbol{w}}$. Here, we go one step farther to consider the **cost-augmented distribution** in Eq. 3. That is, we minimize $\mathbf{KL}(\mathcal{Q}||p'_{\boldsymbol{w}}) := \min_{q \in \mathcal{Q}} \mathbf{KL}(q||p'_{\boldsymbol{w}})$. The next proposition shows that this expression also corresponds to a difference of two log-partition functions, as in Eq. 4.

Proposition 1. *The (regularized) minimization of the cost-augmented KL divergence is equivalent to the following saddle-point problem:*

$$\min_{\boldsymbol{w}} \mathbf{KL}(\mathcal{Q}\|p'_{\boldsymbol{w}}) + \tfrac{\gamma}{2}\|\boldsymbol{w}\|^2 = \min_{\boldsymbol{w}} \max_{\boldsymbol{u} \geq \mathbf{0}} F(\boldsymbol{w}, \boldsymbol{u}) - \boldsymbol{b}^\top \boldsymbol{u} + \tfrac{\gamma}{2}\|\boldsymbol{w}\|^2, \tag{8}$$

where $F(\boldsymbol{w}, \boldsymbol{u}) :=$

$$\sum_{n=1}^{N} \left(\log Z'(\boldsymbol{w}, x^{(n)}) - \log Z'_{\boldsymbol{u}}(\boldsymbol{w}, x^{(n)})\right), \tag{9}$$

with $Z'(\boldsymbol{w}, x)$ *as in Eq. 5, and*

$$Z'_{\boldsymbol{u}}(\boldsymbol{w}, x) := \sum_{y \in \mathcal{Y}(x)} e^{\boldsymbol{w}^\top \boldsymbol{f}(x,y) + \ell(y, \mathcal{Y}(\mathcal{E})) - \boldsymbol{u}^\top \boldsymbol{g}(x,y)}. \tag{10}$$

Proof. See Appendix A. □

In sum, what Proposition 1 shows is that we can easily extend the vanilla PR framework of Ganchev et al. (2010) to incorporate a task-specific cost: by Lagrange duality, the resulting optimization problem still amounts to finding a saddle point of an objective function (Eq. 8), which involves the difference of two log-partition functions (Eq. 9). The difference is that these partition functions now incorporate the cost term $\ell(y, \mathcal{Y}(\mathcal{E}))$. If this cost term has a factorization compatible with the features and the constraints, this comes at no additional computational burden.

[3] Note that the scope of the sum is different in Eqs. 5 and 6: $Z'(\boldsymbol{w}, x)$ sums over *all* coreference trees, while $\widehat{Z}(\boldsymbol{w}, x)$ sums only over those consistent with the gold clusters.

[4] We can always reduce the problem to this case by scaling and adding a constant to the constraint feature vectors.

4.2 Penalized Variant

In their discriminative PR formulation for learning sequence models, Ganchev and Das (2013) optimize an objective similar to Eq. 8 by alternating stochastic gradient updates with respect to $\boldsymbol{w}$ and $\boldsymbol{u}$. In their procedure, $\boldsymbol{b}$ was chosen a priori via linear regression (see their Figure 2).

Here, we propose a different strategy, based on Proposition 1 and a simple observation: while the constraint values $\boldsymbol{b}$ have a more intuitive meaning than the Lagrange multipliers $\boldsymbol{u}$ (since they may correspond, *e.g.*, to proportions of events observed in the data), choosing these upper bounds is often no easier than tuning $\boldsymbol{u}$. In this case, a preferable strategy is to specify $\boldsymbol{u}$ directly—this leaves this variable fixed in Eq. 8, and allows us to get rid of $\boldsymbol{b}$. The resulting problem becomes

$$\min_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{u}) + \tfrac{\gamma}{2}\|\boldsymbol{w}\|^2, \qquad (11)$$

which is a penalized variant of PR and no longer a saddle point problem. This variant requires tuning the Lagrange multipliers u_j in the range $[0, +\infty]$, for every constraint. The two extreme cases of $b_j = 0$ and $b_j = 1$ correspond respectively to $u_j = +\infty$ and $u_j = 0$.[5] Note that this grid search is only appealing for a small number of posterior constraints at corpus-level (since document-level constraints would require tuning separate coefficients for each document).

The practical advantages of the penalized variant over the saddle-point formulation are illustrated in Figure 2, which compares the performance of stochastic gradient algorithms for the two formulations (there, $\eta_2 = 1 - b_2$).

An interesting aspect of this penalized formulation is its resemblance to latent variable models. Indeed, the objective of Eq. 11 is also a **difference of log-partition functions**, as the latent-tree supervised case (cf. Eq. 4). The noticeable difference is that now both partition functions include extra cost terms, either task-specific ($\ell(y, \mathcal{Y}(\mathcal{E}))$ in Z') or with soft constraints ($\boldsymbol{u}^\top \boldsymbol{g}(x, y)$ in $Z'_{\boldsymbol{u}}$). In particular, if we set a single constrained feature $g_1(x, y) := \mathbb{I}(\ell(y, \mathcal{Y}(\mathcal{E})) \neq 0)$ with weight $u_1 \to +\infty$, all non-zero-cost summands in $Z'_{\boldsymbol{u}}(\boldsymbol{w}, x)$ vanish and we get $Z'_{\boldsymbol{u}}(\boldsymbol{w}, x) = \widehat{Z}(\boldsymbol{w}, x)$, recovering the supervised case (see Eq. 6).

[5] This follows from Lagrange duality. If $b_j = 1$, the constraint is vacuous and by complementary slackness we must have $u_j = 0$. If $b_j = 0$, this becomes a hard constraint, so for the nth document, any coreference tree y for which $g_j(x^{(n)}, y) \neq 0$ must have probability zero—this corresponds to setting $u_j \to +\infty$ in Eq. 10.

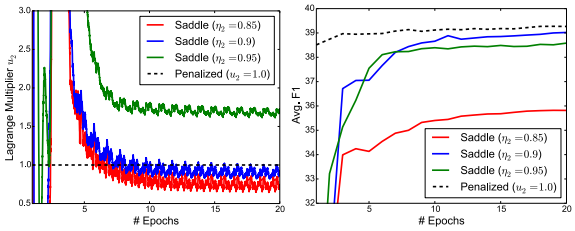

Figure 2: Comparison of saddle-point and penalized PR for Spanish, using the setup in §5.5. Left: variation of the multiplier u_2 over gradient iterations, with strong oscillations in initial epochs and somewhat slow convergence. Right: impact in the averaged F_1 scores (on the dev-set). Contrast with the more "stable" scores achieved by the penalized method.

Intuitively, this formulation pushes probability mass toward structures that respect the constraints in Eq. 7, while moving away from those that have a large task-specific cost. A similar idea, but applied to the generative case, underlies the framework of constrastive estimation (Smith and Eisner, 2005).

4.3 Cost Function

Denote by $\mathcal{E}_m$ the entire coreference chain of the mth mention (so $\mathcal{E} = \bigcup_{m \in \mathcal{M}} \{\mathcal{E}_m\}$), and by $\mathcal{M}_{\text{sing}} := \{m \in \mathcal{M} \mid \mathcal{E}_m = \{m\}\}$ the set of mentions that are projected as singleton in the data (we call this gold-singleton mentions).

We design a task-specific cost $\ell(\widehat{y}, \mathcal{Y}(\mathcal{E}))$ as in Durrett and Klein (2013) to balance three kinds of mistakes: (i) **false anaphora** ($\widehat{y}_m \neq 0$ while $m \in \mathcal{M}_{\text{sing}}$); (ii) **false new** ($\widehat{y}_m = 0$ while $m \notin \mathcal{M}_{\text{sing}}$); and (iii) **wrong link** ($\widehat{y}_m \neq 0$ but $\mathcal{E}_m \neq \mathcal{E}_{\widehat{y}_m}$). Letting $\mathbb{I}_{\text{FA}}(\widehat{y}_m, \mathcal{E})$, $\mathbb{I}_{\text{FN}}(\widehat{y}_m, \mathcal{E})$, and $\mathbb{I}_{\text{WL}}(\widehat{y}_m, \mathcal{E})$ be indicators for these events, we define a weighted Hamming cost function: $\ell(\widehat{y}, \mathcal{Y}(\mathcal{E})) := \sum_{m=1}^{M} (\alpha_{\text{FA}} \mathbb{I}_{\text{FA}}(\widehat{y}_m, \mathcal{E}) + \alpha_{\text{FN}} \mathbb{I}_{\text{FN}}(\widehat{y}_m, \mathcal{E}) + \alpha_{\text{WL}} \mathbb{I}_{\text{WL}}(\widehat{y}_m, \mathcal{E}))$. We set $\alpha_{\text{FA}} = 0.0$, $\alpha_{\text{FN}} = 3.0$, and $\alpha_{\text{WL}} = 1.0$.[6] Since this cost decomposes as a sum over mentions, the computation of cost-augmented marginals (necessary to evaluate the gradient of Eq. 11) can still be done with mention-ranking decoders.

4.4 Constraint Features

Finally, we describe the constraint features (Eq. 7) used in our softmax-margin PR formulation.

Constraint #1: Clusters should not split. Let $|\mathcal{M}| - |\mathcal{E}|$ be the number of anaphoric mentions

[6] The only difference with respect to Durrett and Klein (2013) is that they set $\alpha_{\text{FA}} = 0.1$. We set this coefficient to zero so that all configurations licensed by the constraint features (to be made precise in §4.4) will have zero cost.

in the projected data. We push these mentions to preserve their anaphoricity ($y_m \neq 0$) and to have their antecedent in the projected coreference chain ($\mathcal{E}_m = \mathcal{E}_{y_m}$). To do so, we force the fraction of mentions satisfying these properties to be at least η_1. This can be enforced via a constraint feature

$$g_1(\mathbf{X}, \mathbf{Y}) := \quad (12)$$
$$-\sum_{n=1}^{N} \sum_{m=1}^{M^{(n)}} \mathbb{I}(y_m^{(n)} \neq 0 \ \wedge \ \mathcal{E}_m^{(n)} = \mathcal{E}_{y_m}^{(n)}),$$

and an upper bound $b_1 := -\eta_1 \sum_{n=1}^{N} (|\mathcal{M}^{(n)}| - |\mathcal{E}^{(n)}|)$. (These quantities are summed by a constant and rescaled to meet the assumption in §4.1.) In our experiments, we set $\eta_1 = 1.0$, turning this into a hard constraint. This is equivalent to setting $u_1 = +\infty$ in the penalized formulation.

Constraint #2: Most projected singletons should become non-anaphoric. We define a soft constraint so that a large fraction of the gold-singleton mentions $m \in \mathcal{M}_{\text{sing}}$ satisfy $y_m = 0$. This can be done via a constraint feature

$$g_2(\mathbf{X}, \mathbf{Y}) := \quad (13)$$
$$-\sum_{n=1}^{N} \sum_{m=1}^{M^{(n)}} \mathbb{I}(y_m^{(n)} = 0 \ \wedge \ \mathcal{E}_m^{(n)} = \{m\}),$$

and an upper bound $b_2 := -\eta_2 \sum_{n=1}^{N} |\mathcal{M}_{\text{sing}}^{(n)}|$. In our experiments, we varied η_2 in the range $[0, 1]$, either directly or via the dual variable u_2, as described in §4.1. The extreme case $\eta_2 = 0$ corresponds to a vacuous constraint, while for $\eta_2 = 1$ this becomes a hard constraint which, combined with the previous constraint, recovers bitext direct projection (see §5.3). The intermediate case makes this a **soft constraint** which allows some singletons to be attached to existing entities (therefore introducing some robustness to non-aligned mentions), but penalizes the number of reattachments.

5 Experiments

We now present experiments using the setup in §2. We compare our coreference resolvers trained with softmax-margin PR (§5.5) with three other weakly-supervised baselines: delexicalized transfer with cross-lingual embeddings (§5.2), bitext projection (§5.3), and vanilla PR (§5.4). We also run fully supervised systems (§5.1), to obtain upper bounds for the level of performance we expect to achieve with the weakly-supervised systems.

An important step in coreference resolution systems is **mention prediction**. For English, mention spans were predicted from the noun phrases given by the Berkeley parser (Petrov and Klein, 2007), the same procedure as Durrett and Klein (2013). For Spanish and Portuguese, this prediction relied on the output of the dependency parser, using a simple heuristic: besides pronouns, each maximal span formed by contiguous descendants of a noun becomes a candidate mention. This heuristic is quite effective, as shown by Attardi et al. (2010).

5.1 Supervised Systems

Table 2 shows the performance of supervised systems for English, Spanish and Portuguese. All optimize Eq. 4 appended with an extra regularization term $\frac{\gamma}{2}\|\boldsymbol{w}\|^2$, by running 20 epochs of stochastic gradient descent (SGD; we set $\gamma = 1.0$ and selected the best epoch using the dev-set). All lexicalized systems use the same features as the SURFACE model of Durrett and Klein (2013), plus features for gender and number.[7] We collected a list of pronouns for all languages along with their gender, number, and person information. For English, we trained on the WSJ portion of the OntoNotes dataset, and for Spanish and Portuguese we trained on the monolingual datasets described in §2.

We observe that the Spanish system obtains averaged F_1 scores around 44%, a few points below the English figures.[8] In Portuguese, these scores are significantly lower (in the 37–39% range), which is explained by the fact that the training dataset is much smaller (cf. Table 1).

For English, we also report the performance of **delexicalized** systems, *i.e.*, systems where all the lexical features were removed. The second row of Table 2 shows a drop of 2–2.5 points with respect to the lexicalized system. For the third and fourth rows, the lexical features were replaced by bilingual word embeddings (either English-Spanish or English-Portuguese; a detailed description of these embeddings will be provided in §5.2). Here the drop is small, and for English-Spanish it looks on par with the lexicalized system.

[7] For English, the gender and number of nominal and proper mentions were obtained from the statistics collected by Bergsma and Lin (2006). For Spanish and Portuguese we used a simple heuristic for nominal mentions, based on the determiner preceding the noun (when there is one).

[8] We point out that the supervised Spanish system we present here is strong enough to outperform all participating systems in the SemEval 2010's closed regular track. When trained on the original Spanish SemEval data (with zero- and relative pronoun anaphoras) and evaluated in the provided scorer, it achieves 53.0% averaged F_1 in the test partition; for comparison, TALN-1 (Attardi et al., 2010), the best system at the shared task, achieved 49.6% averaged F_1.

	Dev				Test			
	MUC	B^3	$CEAF_e$	Avg.	MUC	B^3	$CEAF_e$	Avg.
EN lexicalized	**58.35**	**50.75**	**52.08**	**53.73**	**59.07**	**49.25**	48.78	52.37
EN delexicalized, no embed.	56.59	48.81	49.95	51.78	55.96	46.94	46.19	49.70
EN delexicalized, emb. EN-ES	57.55	49.83	51.21	52.86	59.00	**49.25**	**49.00**	**52.42**
EN delexicalized, emb. EN-PT	57.91	49.67	51.01	52.86	58.03	48.16	48.33	51.51
ES lexicalized	**48.24**	**40.97**	**43.59**	**44.27**	**47.03**	**40.68**	**44.09**	**43.93**
PT lexicalized	**35.60**	**34.47**	**42.56**	**37.54**	**41.61**	**36.91**	**40.96**	**39.83**

Table 2: Results for the supervised systems. We show also the performance of delexicalized English systems, with and without cross-lingual embeddings. Shown are MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005), as well their averaged F_1 scores, all computed using the reference implementation of the CoNLL scorer (Pradhan et al., 2014).

	Dev				Test			
	MUC	B^3	$CEAF_e$	Avg.	MUC	B^3	$CEAF_e$	Avg.
ES simple baseline	25.73	24.73	27.89	26.12	26.06	26.12	29.87	27.35
ES baseline #1 (delex. transfer)	33.04	27.47	32.71	31.07	34.35	28.69	34.42	32.49
ES baseline #2 (bitext dir. proj.)	39.42	30.04	38.25	35.90	37.21	29.72	35.97	34.30
ES baseline #3 (vanilla PR)	41.29	33.68	38.56	37.84	39.34	32.95	38.23	36.84
ES softmax-margin PR	**42.34**	**35.53**	**39.95**	**39.27**	**41.22**	**35.30**	**39.94**	**38.82**
PT simple baseline	26.04	26.67	33.19	28.63	22.72	23.91	27.35	24.66
PT baseline #1 (delex. transfer)	22.51	23.27	33.27	26.35	31.11	27.36	32.78	30.42
PT baseline #2 (bitext dir. proj.)	30.43	27.37	36.47	31.42	31.93	27.97	35.40	31.77
PT baseline #3 (vanilla PR)	30.97	27.82	35.14	31.31	**38.39**	33.34	38.73	36.82
PT softmax-margin PR	**33.43**	**31.00**	**38.82**	**34.42**	38.18	**34.05**	**39.47**	**37.23**

Table 3: Results for all the cross-lingual systems. Bold indicates the overall highest scores. As a lower bound, we show a simple deterministic baseline that, for pronominal mentions, selects the closest non-pronominal antecedent, and, for non-pronominal mentions, selects the closest non-pronominal mention that is a superstring of the current mention.

5.2 Baseline #1: Delexicalized Transfer With Cross-Lingual Embeddings

We now turn to the cross-lingual systems. Delexicalized transfer is a popular strategy in NLP (Zeman and Resnik, 2008; McDonald et al., 2011), recently strengthened with cross-lingual word representations (Täckström et al., 2012). The procedure works as follows: a delexicalized model for the source language is trained by eliminating all the language-specific features (such as lexical features); then, this model is used directly in the target language. We report here the performance of this baseline on coreference resolution for Spanish and Portuguese, using the delexicalized models trained on the English data as mentioned in §5.1.

To achieve a unified feature representation, we mapped all language-specific POS tags to universal tags (Petrov et al., 2012). All lexical features were replaced either by **cross-lingual word embeddings** (for words that are not pronouns); or by a universal representation containing the gender, number, and person information of the pronoun. To obtain the cross-lingual word embeddings, we ran the method described by Hermann and Blunsom (2014) for the English-Spanish and English-Portuguese pairs, using the parallel sentences in §2. When used as features, these 128-dimensional continuous representations were scaled by a factor of 0.5 (selected on the dev-set), using the procedure of Turian et al. (2010).

The second and seventh rows in Table 3 show the performance of this baseline, which is rather disappointing. For Spanish, we observe a large drop in performance when going from supervised training to delexicalized transfer (about 11–13% in averaged F_1). For Portuguese, where the supervised system is not so accurate, the difference is less sharp (about 9–11%). These drops are mainly due to the fact that this method does not take into account the intricacies of each language—*e.g.*, possessive forms have different agreement rules in English and in Romance languages;[9] those, on the other hand, have clitic pronouns that are absent in English. Feature weights that promote certain English agreement relations may then harm performance more than they help.

5.3 Baseline #2: Bitext Direct Projection

Another popular strategy for cross-lingual learning is bitext direct projection, which consists in projecting annotations through parallel data in the source and target languages (Yarowsky et al., 2001; Hwa et al., 2005). This is essentially the same as Algorithm 1, except that line 4 is replaced by simple supervised learning, via a minimization

[9] For example, in Figure 1, *their* agrees in number with the possessor (*the alveoli*), but the corresponding *sua* agrees in number and gender with the thing possessed (*função*).

of the loss function in Eq. 4 with ℓ_2-regularization. This procedure has the disadvantage of being very sensitive to annotation errors, as we shall see. For Portuguese, this baseline is a near-reproduction of Souza and Orăsan (2011)'s work, discussed in §6.

The third and eighth rows in Table 3 show that this baseline is stronger than the delexicalized baseline, but still 6–8 points away from the supervised systems. This gap is due to a mix of two factors: prediction errors in the English side of the bitext, and missing alignments. Indeed, when automatic alignments are used, false negatives for coreferent pairs of mentions are common, due to words that have not been aligned with sufficiently high confidence. The direct projection method is not robust to these annotation errors.

5.4 Baseline #3: Vanilla PR

Our last baseline is a vanilla PR approach; this is an adaptation of the procedure carried out by Ganchev and Das (2013) to our coreference resolution problem. The motivation is to increase the robustness of bitext projection to annotation errors, which we do by applying the soft constraints in §4.4. We seek a saddle-point of the PR objective by running 20 epochs of SGD, alternating $\boldsymbol{w}$-updates and $\boldsymbol{u}$-updates. The best results in the dev-set were obtained with $\eta_1 = 1.0$ and $\eta_2 = 0.9$.

By looking at the fourth and ninth rows of Table 3, we observe that vanilla PR manages to reduce the gap to supervised systems, obtaining consistent gains over the bitext projection baseline (with the exception of the Portuguese dev-set). This confirms the ability of PR methods to handle annotation mistakes in a robust manner.

5.5 Our Proposal: Softmax-Margin PR

Finally, the fifth and last rows in Table 3 show the performance of our systems trained with softmax-margin PR, as described in §4.1. We optimized the loss function in Eq. 11 with $\gamma = 1.0$ by running 20 epochs of SGD, setting $u_1 = +\infty$ and $u_2 = 1.0$ (cf. §4.4)—the last value was tuned in the dev-set. As shown in Figure 2, this penalized variant was more effective than the saddle point formulation.

From Table 3, we observe that softmax-margin PR consistently beats all the baselines, narrowing the gap with respect to supervised systems to about 5 points for Spanish, and 2–3 points for Portuguese. Gains over the vanilla PR procedure (the strongest baseline) lie in the range 0.5–3%. These gains come from the ability of softmax-margin PR to handle task-specific cost functions, enabling a better management of precision/recall tradeoffs.

5.6 Error Analysis

We carried out some error analysis, focused on the Spanish development dataset, to better understand where the improvements of softmax-margin PR come from. The main conclusions carry out to the Portuguese case, with a few exceptions, mostly due to different human annotation criteria.

Table 4 shows the precision and recall scores for mention prediction and the different coreference evaluation metrics. Note that all systems predict the same candidate mentions; however a final post-processing discards all mentions that ended up in singleton entities, for compliance with the official scorer. Therefore, the mention prediction score reflects how well a system does in predicting if a mention is anaphoric or not. The first thing to note is that the PR methods, due to their ability to create new links during training (via constraint #2) tend to predict fewer singletons than the direct projection method. Indeed, we observe that soft max-margin PR achieves 47.1% mention prediction recall, which is more than 5% above the direct projection method, and 10% above the delexicalized transfer method. Note also that, while the vanilla PR method achieves higher recall than the two other baselines, it is still almost 5% below the system trained with soft-max margin PR. This is because vanilla PR does not benefit from the cost function in §4.3—such cost is able to penalize false non-anaphoric mentions and encourage larger clusters, allowing softmax-margin PR to achieve a better precision-recall trade-off. From Table 4, we can see that this improvement in mention recall consistently translates into higher recall for the MUC, B^3 and $CEAF_e$ coreference metrics.

Further analysis revealed that a major source of error for the delexicalized baseline is its inability to handle pronominal mentions robustly across languages—as hinted in footnote 9. In practice, we found the delexicalized systems to be quite conservative with possessive pronouns: for the Spanish dataset, where the vast majority of possessive pronouns are anaphoric, the delexicalized model incorrectly predicts 53.3% of these pronouns as non-anaphoric. The direct projection model is slightly less conservative, missing 30.1% of the possessives (arguably due to its inability to recover missing links in the projected data, dur-

	Mention	MUC	B^3	$CEAF_e$
delex.	37.1 / 62.2	25.6 / 46.5	19.6 / 45.7	27.9 / 39.5
dir. proj.	41.7 / 77.5	29.3 / 60.4	19.2 / 69.5	31.8 / 47.9
vanilla PR	42.2 / 78.0	30.9 / 62.3	23.2 / 61.7	31.9 / 48.8
our PR	47.1 / 74.1	33.7 / 57.1	26.0 / 56.1	34.9 / 46.7

Table 4: Recall/precision scores for mention prediction, MUC, B^3 and $CEAF_e$, all computed in the Spanish dev set.

ing training). By comparison, the vanilla and softmax margin PR models only miss 4.9% and 3.4% of the possessives, respectively. In Portuguese, where many possessives are not annotated in the gold data, we observe a similar but much less pronounced trend.

6 Related Work

While multilingual coreference resolution has been the subject of recent SemEval and CoNLL shared tasks, no submitted system attempted cross-lingual training. As shown by Recasens and Hovy (2010), language-specific issues pose a challenge, due to phenomena as pronoun dropping and grammatical gender that are absent in English but exist in other languages. We have discussed some of these issues in the scope of the present work.

Harabagiu and Maiorano (2000) and Postolache et al. (2006) projected English corpora to Romanian to bootstrap human annotation, either manually or via automatic alignments. Rahman and Ng (2012) applied translation-based projection at test time (but require an external translation service). Hardmeier et al. (2013) addressed the related task of cross-lingual pronoun prediction. While all these approaches help alleviate the corpus annotation bottleneck, none resulted in a full coreference resolver, which our work accomplished.

The work most related with ours is Souza and Orăsan (2011), who also used parallel data to transfer an English coreference resolver to Portuguese, but could not beat a simple baseline that clusters together mentions with the same head. Their approach is similar to our bitext direct projection baseline, except that they used Reconcile (Stoyanov et al., 2010) instead of the Berkeley Coreference System, and a smaller version of the FAPESP corpus. We have shown that our softmax-margin PR procedure is superior to this approach.

Discriminative PR has been proposed by Ganchev et al. (2010). The same idea underlies the generalized expectation criterion (Mann and McCallum, 2010; Wang and Manning, 2014). An SGD algorithm for solving the resulting saddle point problem has been proposed by Liang et al. (2009), and used by Ganchev and Das (2013) for cross-lingual learning of sequence models. We extended this framework in two aspects: by incorporating a task-specific cost in the objective function, and by formulating a penalized variant of PR.

7 Conclusions

We presented a framework for cross-lingual transfer of coreference resolvers. Our method uses word-aligned bitext to project information from the source to the target language. Robustness to projection errors was achieved via a PR framework, which we generalized to handle task-specific costs, yielding softmax-margin PR. We also proposed a penalized formulation that is effective for a small number of corpus-based constraints. Empirical gains were shown over three popular cross-lingual methods: delexicalized transfer, bitext direct projection, and vanilla PR.

Acknowledgments

I would like to thank the reviewers for their helpful comments, José Guilherme Camargo de Souza for pointing to existing datasets, and Mariana Almeida for valuable feedback. This work was partially supported by the EU/FEDER programme, QREN/POR Lisboa (Portugal), under the Intelligo project (contract 2012/24803), and by the FCT grants UID/EEA/50008/2013 and PTDC/EEI-SII/2312/2012.

A Proof of Proposition 1

Let us fix $\boldsymbol{w}$ and see how to evaluate $\mathbf{KL}(\mathcal{Q}||p'_{\boldsymbol{w}}) = \min_{q\in\mathcal{Q}} \mathbf{KL}(q||p'_{\boldsymbol{w}})$. We have:

$$\begin{aligned}\mathbf{KL}(q\|p'_{\boldsymbol{w}}) &= -\mathbf{H}(q) - \textstyle\sum_{\mathbf{Y}} q(\mathbf{Y}) \log p'_{\boldsymbol{w}}(\mathbf{Y}|\mathbf{X}) \\ &= -\mathbf{H}(q) + \textstyle\sum_n \log Z'(\boldsymbol{w}, x^{(n)}) - \\ &\quad \textstyle\sum_{\mathbf{Y}} q(\mathbf{Y})(\boldsymbol{w}^\top \boldsymbol{f}(\mathbf{X},\mathbf{Y}) + \ell(\mathbf{Y})),\end{aligned}$$

where $\ell(\mathbf{Y}) := \sum_{n=1}^N \ell(y, \mathcal{Y}(\mathcal{E}^{(n)}))$ and $\boldsymbol{f}(\mathbf{X},\mathbf{Y}) := \sum_{n=1}^N \boldsymbol{f}(x^{(n)}, y^{(n)})$. Introducing Lagrange multipliers $\boldsymbol{u}$ for the posterior constraints, we get the Lagrangian function:

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{u}) &= -\mathbf{H}(q) + \textstyle\sum_n \log Z'(\boldsymbol{w}, x^{(n)}) - \boldsymbol{b}^\top \boldsymbol{u} \\ &\quad - \textstyle\sum_{\mathbf{Y}} q(\mathbf{Y})(\boldsymbol{w}^\top \boldsymbol{f}(\mathbf{X},\mathbf{Y}) + \ell(\mathbf{Y}) - \boldsymbol{u}^\top \boldsymbol{g}(\mathbf{X},\mathbf{Y})).\end{aligned}$$

By standard variational arguments (namely, Fenchel duality between the the log-partition function and the negative entropy; see *e.g.* Martins et al. (2010)), we have that the optimal q^* that minimizes the Lagrangian is

$$q^*(\mathbf{Y}) = \frac{e^{\boldsymbol{w}^\top \boldsymbol{f}(\mathbf{X},\mathbf{Y}) + \ell(\mathbf{Y}) - \boldsymbol{u}^\top \boldsymbol{g}(\mathbf{X},\mathbf{Y})}}{\prod_{n=1}^N Z'_{\boldsymbol{u}}(\boldsymbol{w}, x^{(n)})}.$$

Plugging this in the Lagrangian yields Eq. 8.

References

Mariana S. C. Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. 2015. Aligning opinions: Cross-lingual opinion mining with dependencies. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. TANL-1: coreference resolution by parse analysis and similarity clustering. In *Proc. of the International Workshop on Semantic Evaluation*.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proc. of International Conference on Language Resources and Evaluation: Workshop on Linguistics Coreference*.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proc. of Empirical Methods in Natural Language Processing*.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Sandra Collovini, Thiago Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In *Workshop em Tecnologia da Informação e da Linguagem Humana*.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proc. of Empirical Methods in Natural Language Processing*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proc. of Empirical Methods in Natural Language Processing*.

Greg Durrett, David Hall, and Dan Klein. 2013. Decentralized entity-level modeling for coreference resolution. In *Proc. of Annual Meeting of the Association for Computational Linguistics*.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 41–48.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proc. of Empirical Methods in Natural Language Processing*.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-Margin CRFs: Training Log-Linear Models with Loss Functions. In *NAACL*.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proc. of Annual Meeting of the Association for Computational Linguistics*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Sanda M Harabagiu and Steven J Maiorano. 2000. Multilingual coreference resolution. In *Proc. of the Conference on Applied Natural Language Processing*.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proc. of Empirical Methods in Natural Language Processing*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proc. of North American Chapter of the Association of Computational Linguistics*.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proc. of International Conference on Machine Learning*, pages 641–648.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gideon Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984.

André F. T Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo Parsers: Dependency Parsing by Approximate Variational Inference. In *Proc. of Empirical Methods for Natural Language Processing*.

André F. T Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proc. of Empirical Methods in Natural Language Processing*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of the Annual Meeting on Association for Computational Linguistics*.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proc. of Empirical Methods in Natural Language Processing*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proc. of Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 404–411.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. of LREC*.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proc. of the International Conference on Language Resources and Evaluation*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proc. of the Conference on Computational Natural Language Learning: Shared Task*.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proc. of Empirical Methods in Natural Language Processing*.

Altaf Rahman and Vincent Ng. 2011. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research*, 40(1):469–521.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44(4):315–345.

Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. of the International Workshop on Semantic Evaluation*.

Noah A Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proc. of Annual Meeting on Association for Computational Linguistics*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

José Guilherme Camargo de Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Anaphora Processing and Applications*, pages 59–69. Springer.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with reconcile. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. of the North American Chapter of the Association for Computational Linguistics*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Ivan Titov and Alexandre Klementiev. 2012. Cross-lingual induction of semantic roles. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Proc. of the Annual Meeting of the Association for Computational Linguistics: Demo Session*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the Conference on Message Understanding*, pages 45–52. Association for Computational Linguistics.

Mengqiu Wang and Chris Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. of the First International Conference on Human Language Technology Research*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.