Named Entity Recognition for Linguistic Rapid Response in Low-Resource Languages: Sorani Kurdish and Tajik

Patrick Littell, Kartik Goyal, David Mortensen, Alexa Little, Chris Dyer, Lori Levin
Carnegie Mellon University
Language Technologies Institute
5000 Forbes Ave., Pittsburgh PA 15213
plittell@cs.cmu.edu, kartikgo@cs.cmu.edu, dmortens@cs.cmu.edu
alexanicolelittle@gmail.com, cdyer@cs.cmu.edu, lsl@cs.cmu.edu

Abstract

This paper describes our construction of named-entity recognition (NER) systems in two Western Iranian languages, Sorani Kurdish and Tajik, as a part of a pilot study of *Linguistic Rapid Response* to potential emergency humanitarian relief situations. In the absence of large annotated corpora, parallel corpora, treebanks, bilingual lexica, etc., we found the following to be effective: exploiting distributional regularities in monolingual data, projecting information across closely related languages, and utilizing human linguist judgments. We show promising results on both a four-month exercise in Sorani and a two-day exercise in Tajik, achieved with minimal annotation costs.

1 Introduction

This paper describes our rapid construction of NER systems, as a part of a pilot study of *Linguistic Rapid Response* to potential emergency humanitarian relief situations. When a disaster strikes a community that speaks a low-resource language without an existing NLP infrastructure, how long would it take to put such an infrastructure, however imperfect, in place? What kinds of systems can be in place within 24 or 48 hours, and what levels of performance can we expect? What resources – besides existing gazetteers, parallel corpora, etc. – can be assembled in this timeframe and brought to bear on NER?

Contemporary techniques for creating natural language processing (NLP) tools are dominated by supervised learning approaches, in which large quantities of high quality data are annotated in a task-specific fashion and utilized along with manually-built collaborate resources like WordNet, Ontonotes, etc. These techniques have provided very good performance, but with 7,000 languages in the world these resources cannot feasibly be compiled in advance, and could not be assembled within an emergency timeframe.

Following many examples of "surprise language exercises" (Oard, 2003), the work described in this paper tackles the problem of rapid development of important NLP tools and resources, with a focus on NER, for low resource languages. We assume that for a "low resource" language, there is some monolingual corpus data and little to no annotated data for supervised training. A major theme underlying this work is a focus on building "omnivorous" models and pipelines that, instead of relying on elaborate and robust linguistic resources compiled ahead of time, try to opportunistically incorporate linguistic theory, informal and non-expert intuitions about the language and task at hand, and resources adapted from closely related languages, all while avoiding extensive manual annotation.

An important feature of our work is the use of human linguists who do not speak the language but are familiar with the structure of human languages in general, have some knowledge about the language family in question, and can absorb facts about the language quickly by reading reference grammars and looking at data. Specifically, in this project, linguists worked interactively with unsupervised morphology induction, annotated named entities, and identified thresholds for automatic tagging of multi-word named entities.

We illustrate this approach for two low resource languages, Sorani Kurdish and Tajik, provided during two surprise language evaluations. Software development and Sorani NER processing took place over

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 998–1006, Osaka, Japan, December 11-17 2016.

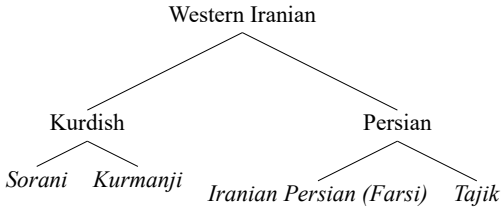

Figure 1: Sorani, Tajik, and selected relatives

a four-month period, while the Tajik surprise-language event took place in only 36 hours. Hence, all techniques described in this paper, whether performed by machines or humans or both, have runtimes in minutes or hours rather than days.

2 Data sources

2.1 Sorani Kurdish

Sorani (or "Central") Kurdish is a Western Iranian language in the Kurdish family, spoken by about 6.7 million people in Iraqi Kurdistan and the Kurdistan Province of Iran. Unlike other Kurdish languages, but like the more distantly related Persian (or "Farsi"), it is written in a Perso-Arabic script, albeit with modifications (like additional vowel glyphs) that make it more suitable for writing Sorani. We obtained Sorani monolingual data from the LCTL language pack (Simpson et al., 2008).

2.2 Kurmanji Kurdish

Some of the challenge of Sorani NER is orthographic in nature, since its Perso-Arabic script, while closer to the phonemic form of words than most other Perso-Arabic scripts, still has some significant ambiguities in vowel representation, and lacks an uppercase-lowercase distinction, which is an important feature for NER. In order to partially mitigate these ambiguities, we also made use of Kurmanji (or "Northern") Kurdish data from the Pewan news corpus (Esmaili et al., 2013). Kurmanji is spoken primarily in eastern Turkey by about 20 million speakers, and unlike Sorani is written in a Roman script. Sorani and Kurmanji are sometimes described as dialects of the same languages, but sometimes described as different languages due to their significant morphological differences.

2.3 Tajik Persian

Tajik is a variety of Persian spoken primarily in Tajikistan by about 8 million speakers, and is written primarily in Cyrillic script. We obtained Tajik monolingual data from the Leipzig corpus of news crawls (Biemann et al., 2007).

2.4 IPA representations

So that resources in different languages and scripts could be more directly compared, and so that judgments about the data could be made rapidly by linguists without native proficiency in the Perso-Arabic and Cyrillic writing systems, we produced representations of the Sorani, Kurmanji, and Tajik data in the International Phonetic Alphabet (IPA). To disambiguate ambiguous Sorani forms, we used a conditional random field (CRF) (Lafferty et al., 2001) that utilized a combination of human judgments, universal phonetic features, and language models of related languages; we describe this "IPAization" process in more detail in Mortensen et al. (2016) and Littell et al. (2016).

3 Named Entity Recognition

For the NER task, we focused on identifying mentions of persons (PER), locations (LOC), and organizations (ORG) in the textual data. Our core system is a CRF based system with L1-regularization, where x is the input sequence and output y is the appropriate tag sequence, and no features look beyond a history of length 1.

$$f(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} f(\mathbf{x}, \mathbf{y}_i, \mathbf{y}_{i-1}). \quad (1)$$

Based on previous work in the area (Tjong Kim Sang and De Meulder, 2003; McCallum and Li, 2003; Sha and Pereira, 2003), we begin with a standard set of features commonly used for training NER systems:

- Current token and Current tag
- Previous token and Current tag
- Next token and Current tag
- Current+previous token and current tag
- Current+next token and current tag
- First five features but with previous tag
- First five features conjoined with both current and previous tags
- Contains foreign script characters
- Indicator features for tokens containing digits
- Features about capitalization information
- Prefix features

However, given the paucity of training data, these features are numerous and sparse, we do not expect the CRF model to perform well with standard features alone, even with L1 regularization.

Hence, keeping up with the general theme of this work, we aim to induce features in an unsupervised manner by exploiting the distributional regularities in the monolingual data, guiding the unsupervised sub-components to encode our intuitions and knowledge about the task and the target language, and utilize features from closely-related languages.

3.1 Gazetteers

For Sorani, the absence of a capitalization feature in its Perso-Arabic script posed a difficulty, as capitalization serves as a valuable feature for NER. However, as noted in §2, the closely related Kurmanji Kurdish is written in a Roman script that does distinguish capitalization.

Using the IPA representations of the Sorani and Kurmanji texts and a frequency-weighted edit distance algorithm, we inferred for each Sorani token a corresponding Kurmanji token (e.g., ⟨ɛvɣanistan,ɛfganistan⟩), and assigned to each Sorani token the capitalization frequency of the Kurmanji word (in this case, *Efganistan*, which had a capitalization frequency of 1.0). These features were used a "probabilistic gazetteer" in lieu of a real Sorani gazetteer, where the capitalization frequency is treated as if it represented the probability that a word occurred in a gazetteer. We describe this gazetteer inference strategy in greater detail in Littell et al. (2016).

For Tajik, we constructed a more traditional gazetteer using Tajik's relatively extensive Wikipedia. Since Wikipedia titles are linked between Wikipedias in different languages, we had parallel English and Tajik titles; we filtered the English titles by heuristics including capitalization, and used the corresponding Tajik titles as the gazetteer entries.

3.2 Unsupervised morphology induction

The Western Iranian languages are morphologically rich, with Sorani in particular having a high degree of morphological complexity (Walther, 2011; Esmaili and Salavati, 2013). In such languages, the presence of certain morphemes is strongly correlated with certain grammatical functions being present, which could be informative for the problem of identifying and discriminating named entities.[1]

While unsupervised induction of morphological grammars is a long-standing problem, the inferred morphological analyses typically diverge from conventional linguistic analyses rather substantially. We addressed this shortcoming by using feedback from human linguists using a modification of the interactive learning paradigm proposed by Hu et al. (2011). While superficially related to active learning (Settles,

[1] While Western Iranian languages also utilize prefixes and root modification, we concentrated here on suffixes alone; this simplifies the model and concentrates on those morphological alternations we believe more likely to be relevant to NER.

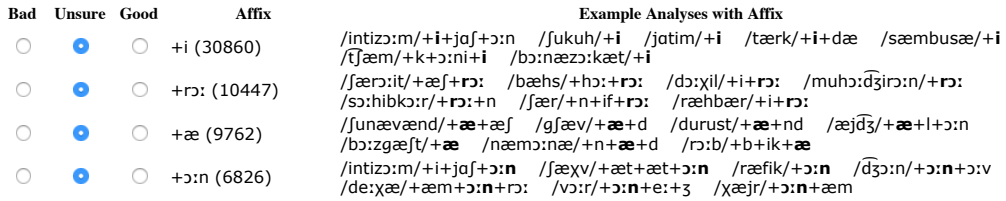

Figure 2: Screenshot of the feedback interface containing analyses for Tajik.

2012), the interactive paradigm charges the annotator with the task of identifying subjective systematic errors, instead of picking new instances to be annotated. In case of morphology learning, we let linguists give their judgment on the quality of affixes, which is used to constrain the hypothesis space and tune the parameters of the unsupervised learner. The feedback interface is shown in Fig. 2.

We used a hierarchical Bayesian model of morphological segmentation. Our prior expectations are (1) that stems should be more diverse than suffixes, but both should be reused when possible, (2) that individual suffixes are likely to be very short, and (3) that suffixes are likely to proceed in a characteristic order (e.g., *-ion* and *-al* are both English suffixes, but *-ion-al* is valid and occurs in many words, for example in *inspirational*, while *-al-ion* is not valid).

To encode the first two assumptions, we assumed that the distributions over stems and suffixes are governed by a Dirichlet processes with parameters set to encourage lower entropy samples for suffixes and higher entropy samples for stems. The base distribution $\mathsf{Word}(\lambda)$ is a process that generates a word by sampling a length from a Poisson distribution with mean λ and then choosing characters randomly for each position. To capture the fact that affixes proceed in a characteristic order, we in turn assumed these were generated by a bigram Markov process governed by a hierarchical Dirichlet process (Teh et al., 2006). The generative process is stated in Algorithm 1.

Algorithm 1 Morphological induction

$\boldsymbol{\theta} \sim \mathsf{DP}(\alpha_1, \mathsf{Word}(\lambda = 6))$
$\boldsymbol{\varphi} \sim \mathsf{DP}(\alpha_2, \mathsf{Word}(\lambda = 1))$
$\boldsymbol{\varphi}_{\cdot|x} \sim \mathsf{DP}(\alpha_3, \boldsymbol{\varphi})\ \forall x \in \Sigma^*$
for `each word` w `in surface vocabulary` V **do**
  `Draw # of suffixes` $\ell \sim \mathsf{Geom}(\rho = 0.9)$
  `Draw stem` $b \sim \mathsf{Cat}(\boldsymbol{\theta})$
  $s_{-1} = \langle \mathsf{b} \rangle$
  for `each suffix index` i `from 1 to` ℓ **do**
    `Draw affix` $s_i \sim \mathsf{Cat}(\boldsymbol{\varphi}_{\cdot|s_{-1}})$
    $w = w + s_i$
    $s_{-1} = s_i$
  end for
end for

Since our model does not depend on anything but word types, the observed data is just the vocabulary (in IPA form) of the target language, and the goal of inference is to find the distribution over segmentations given our model and the vocabulary. To do so, we use block Gibbs sampling (marginalizing the draws from the Dirichlet processes). Since we are considering all analyses of a word at once, constraints against certain morphemes are trivial to incorporate. For the experiments reported below, we ran 1000 iterations of Gibbs sampling, then obtained feedback followed by a further 1000 iterations twice.

3.3 Unsupervised class induction: Hard clustering

One way of reducing sparseness due to the lexicalization of the features is to map the types or tokens of the monolingual corpus to a smaller number of classes. We try to obtain mappings such that the tokens/types sharing similar characteristics are mapped to one class. For example, if "Lord" and "Lady" are mapped to the same class, then two different sequences "Lord Palmerston" and "Lady Grey" will share the class information and will, if "Palmerston" is annotated as a name in the training data, be more likely to predict "Grey" as a name as well. We focus on context-aware class induction, particularly modeling the classes

with a first-order Markovian assumption.

Brown et al. (1992) introduced a bottom-up agglomerative word clustering algorithm which generates a hard clustering (i.e., a word belongs to only one cluster). With this hard clustering assumption, we aim to achieve a clustering C that maximizes the log-likelihood of the data, $\log P(w_1, ...w_n, \mathbf{C}(w_1), ..\mathbf{C}(w_n))$, i.e.

$$\arg\max_{\mathbf{C}} \log \prod_{i=1}^{n} p(\mathbf{C}(w_i) \mid \mathbf{C}(w_{i-1})) \times p(w_i \mid \mathbf{C}(w_i))$$

We experimented with 500 and 1000 clusters. Manual qualitative analysis of these clusters revealed that they created several meaningful groups: foreign words, numbers, names, etc.

3.4 Unsupervised class induction: Soft clustering with the influence of collocations

In addition to the traditional "hard" Brown clusterings, we also experimented with soft clusterings where the parameters can be influenced by external knowledge, using Expectation Maximization (Dempster et al., 1977). We focused on influencing our NER system using information about collocations – bigrams that co-occur with frequency greater than chance – in the monolingual text. The intuition behind collocations is that many names of people ("barak obama"), places ("arabistani saudi"), and organizations ("bomdodi telefonhoi") are expected to be identified as collocations.

We warm-started with the distributions obtained by the Brown cluster algorithms, smoothed via additive (dirichlet) smoothing. Our pilot experiments showed that this resulted in better performance when compared to the performance with random initializations. The runtime (O($token * iterations * C^2$)) is higher than that of Brown algorithm (O($types * C^2 + tokens$)) because we estimate all the distributions of the probabilistic HMM using dynamic programming (Rabiner and Juang, 1986). Hence, subsequent models used interpolated stochastic batch updates (Liang and Klein, 2009) instead of batch updates so that the convergence is faster.

We used a likelihood ratio test (Dunning, 1993), which is a form of hypothesis testing that decides whether the second word in the bigram is unusually associated with the first word of the bigram or not, to determine an initial list of possible NE collocations. For both Sorani and Tajik, this measure results in desirable LR score graphs for bigrams with a distinct "elbow" for both the languages. However, determining the threshold, below which a collocation should not be considered genuine for the purposes of further steps, was done manually by a human linguist looking at IPA representations of the collocations. This is a judgment that need not be made by a native speaker, or even an expert in the language in question, but just someone with some familiarity with the language or language group, the ability to read IPA, and enough real-world knowledge to recognize when a list of two-word phrases in IPA switches from mostly referring to names and places, to referring to names and places only occasionally.

We use these collocations, as generated by the likelihood ratio test and thresholded by human judgment, to bias unsupervised class induction over words in the target language, so that collocations are encouraged to fall into the same clusters. To our knowledge, this work is the first to bias unsupervised class induction using collocation knowledge.[2]

EM based optimization allows us to bias the parameters of the HMM, to encourage collocations to fall into same clusters. Hence, whenever we observe collocations in the monolingual data during the E step, we use an Identity matrix as the transition matrix, i.e. $P(C(w_i) \mid C(w_{i-1})) = 1$. The M step is performed as usual, resulting in learning of parameters affected by the biased expectation counts.

As discussed above, constraining the collocation members to belong to the same clusters is attractive, but the collocations that we estimated automatically are certainly not pure. Hence, we introduce posterior regularization (PR) (Ganchev et al., 2010) into our HMM inference algorithm. We want the posterior of the HMM distribution to reflect the fact that adjacent tokens in identified collocations in the monolingual data tended to belong to same clusters.

[2]It should be noted that this is different from work in Liang (2005), which used the mutual information between adjacent types directly as features in the learning model. We avoid this method to keep our feature space small, in light of the paucity of training data.

For this technique, if we denote the original HMM distribution by $p(\mathbf{C} \mid \mathbf{W})$ with parameters θ, and a variational approximation $q(\mathbf{C})$ to the original distribution which respects our collocation based constraints i.e. $E_q(\phi(\mathbf{W}, \mathbf{C})) = -1$ where, $\phi(w_i, C(w_i), C(w_{i-1})) = -1$ if $C(w_i) = C(w_{i-1})$ and 0 otherwise, for $w_i s$ that are members of collocations; then the objective that we optimize becomes:

$$\arg\min_{\theta} \text{KL}(q \mid\mid p) \text{subject to } E_q(\phi(\mathbf{W}, \mathbf{C})) = -1 + \epsilon$$

When this objective is solved using its dual, the variational approximation q (after optimization for the dual variable λ) looks like:

$$q^*(\mathbf{C}) = \frac{p_\theta(\mathbf{C}|\mathbf{W}) exp(-\lambda^*.\phi(\mathbf{W}, \mathbf{C}))}{Z(\lambda^*)}$$

Since the constraints ϕ are local at the level of transition probabilities, the PR solution can be easily incorporated into the dynamic program of HMM.

3.5 Experiments

We present our results on the task of named entity recognition for Sorani and Tajik. For Sorani, the training (2175 instances) and test (212 instances) data was obtained from the annotated NER data in the LCTL language pack. For Tajik, we had access to a native speaker who, in about four hours, annotated 600 examples, which we split into 250 training instances, 250 test instances, and 100 development instances.

Features	**Rec.**	**Prec.**	**F1**
Std.	0.412	0.694	0.517
Std.+Br	0.476	0.702	0.567
Std.+Br+Gaz	0.490	0.750	0.593
Std.+Br+Gaz+Mph	0.509	**0.751**	**0.606**
Std.+PR+Gaz+Mph	**0.513**	0.741	**0.606**

Table 1: Results on Sorani NER

Features	**Rec.**	**Prec.**	**F1**
Std	0.302	0.709	0.423
Std+Br	0.511	0.657	0.574
Std+Br+Gaz	0.512	0.656	0.575
Std+Br+Gaz+Mph	0.517	**0.668**	**0.583**
Std+PR+Gaz+Mph	**0.537**	0.637	**0.583**

Table 2: Results on Tajik NER

In Tables 1 and 2, 'Std' refers to the standard features used in supervised NER systems (§3), 'Br' to the class features obtained from the Brown algorithm (§3.3), 'PR' to the class features from the EM and posterior regularization algorithm (§3.4), 'Gaz' to Gazetteer based features (§3.1), and 'Mph' to features from morphology induction (§3.2).

As we can observe, systems based only on standard features (§3) perform comparatively poorly, while adding Brown clusters lead to a large gain in recall especially in the Tajik condition. In both languages, the Brown and PR conditions perform similarly on F1, with the Brown condition having higher precision and the PR condition having higher recall. The reduction of precision in the PR condition is most likely a result of expanding the feature space with both Brown clusters and EM-based clusters.

The "gazetteer" for Sorani, which attempts to fabricate capitalization values for Sorani by comparison with Kurmanji words (§3.1), led to improvements in both recall and precision, while the Tajik gazetteer, collected from Tajik Wikipedia titles, did not lead to significant gains. Manual inspection of the resulting

	Sorani			Tajik		
Features	**Rec.**	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
PER	0.406	0.709	0.516	0.446	0.458	0.452
LOC	0.680	0.801	0.735	0.589	0.737	0.655
ORG	0.343	0.604	0.438	0.136	0.375	0.200

Table 3: Error analysis on NER

Features	**Rec.**	**Prec.**	**F1**
Std+PR+Gaz+Mph	0.409	0.669	0.508

Table 4: Results on Tajik NER, using 1350 linguist annotations in lieu of native-speaker annotations

gazetteer revealed it to be fairly noisy with respect to the forms of the names; while the entries appeared to be, for the most part, genuinely named entities, many of them appear to have been converted directly from the Persian Wikipedia. Since Persian script does not completely represent vowels, the Tajik authors in many cases were likely guessing at the vowels when they were unfamiliar with the named entity.

Morphological features (§3.2) provided small improvements for both languages. Interestingly, the morphological features affected both the languages differently; while the Sorani system relied more on the induced stems, the Tajik system relied more on the suffixes. This may reflect the complexity of Sorani morphology (Walther, 2011; Esmaili and Salavati, 2013), in which many apparent affixes are actually enclitic, and therefore might not provide category information about their hosts as reliably as true suffixes do.

In Table 3, we observe that our NER systems are best at identifying LOC and are slightly worse at identifying PER. However, they perform substantially worse on identifying ORG because their proper noun parts can be confused with both locations and persons, and they often involve common words (e.g. "association", "for", etc.) that in other contexts are not part of NEs. Note that unlike in English, capitalization is not as helpful in distinguishing multiword NEs in Sorani and Tajik. Sorani "capitalization" here is only a feature inferred on a word-by-word basis from Kurmanji text, as detailed in §3.1, and Tajik generally uses Russian-style capitalization conventions in which only the first word in a multiword NE needs to be capitalized, making ORG identification much more difficult than in languages that capitalize all or most words in an ORG.

For the Tajik condition, we also had linguists – without prior experience in Tajik but generally familiar with Western Iranian languages – annotate another 1350 instances in sixteen person-hours. This was made possible by the IPA conversion step mentioned in §2.4, since the linguists did not have native proficiency in reading Cyrillic text.

Using this larger set instead of the smaller native-speaker-annotated set, we achieved similar results, with lower recall but higher precision (Table 4). This is a promising result, as it suggests that a team of linguists, even those without prior familiarity with the language, can create useful training data in a short time even when native informants are unavailable.

4 Conclusion

Our work demonstrates that, by using tools, data resources, and human resources (like linguists and language consultants) in innovative ways, it is possible to overcome some of the obstacles to developing standard NLP tools like NER systems for low-resource languages.

We built Named Entity Recognizers for Sorani Kurdish and Tajik in a manner which, while requiring minimal human annotator effort, managed to successfully incorporate informal intuitions and linguistic knowledge about the task and the languages into the system and seeks to identify and exploit latent patterns in the monolingual data.

To this end, we developed unsupervised class induction systems that were influenced by noisy collocation lists, and morphology induction systems that could be biased by subjective human feedback.

Moreover, we also showed that mapping the orthographic representation of a language to a general phonological representation not only enables efficient human analysis and annotation, but also opens avenues for transferring linguistic information from related languages.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office, under the LRRT extension to contract/grant number W911NF10-1-0533.

References

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection: Monolingual corpora of standard size. *Proceedings of Corpus Linguistic 2007*.

Peter F. Brown, Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Kyumars Sheykh Esmaili and Shahin Salavati. 2013. Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 300–305.

Kyumars Sheykh Esmaili, Shahin Salavati, Somayeh Yosefi, Donya Eliassi, Purya Aliabadi, Shownem Hakimi, and Asrin Mohammadi. 2013. Building a test collection for Sorani Kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7.

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

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *Proc. ACL*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Percy Liang and Dan Klein. 2009. Online EM for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619. Association for Computational Linguistics.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Patrick Littell, David Mortensen, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Bridge-language capitalization inference in Western Iranian: Sorani, Kurmanji, Zazaki, and Tajik. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*, pages 3318–3324.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

David Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Douglas W. Oard. 2003. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):79–84, September.

Lawrence Rabiner and Biing-Hwang Juang. 1986. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Burr Settles. 2012. *Active Learning*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. *Collaboration: Interoperability between people in the creation of language resources for less-resourced languages*, page 7.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, vol. 4*, pages 142–147. Association for Computational Linguistics.

Géraldine Walther. 2011. Fitting into morphological structure: Accounting for Sorani Kurdish endoclitics. In A. Ralli, G. Booij, S. Scalise, and A. Karasimos, editors, *Proceedings of the 8th Mediterranean Morphology Meeting*, pages 299–321.