Accurate Unsupervised Joint Named-Entity Extraction from Unaligned Parallel Text

Robert Munro
Department of Linguistics
Stanford University
Stanford, CA 94305
rmunro@stanford.edu

Christopher D. Manning
Department of Computer Science
Stanford University
Stanford, CA 94305
manning@stanford.edu

Abstract

We present a new approach to named-entity recognition that jointly learns to identify named-entities in parallel text. The system generates seed candidates through local, cross-language edit likelihood and then bootstraps to make broad predictions across both languages, optimizing combined contextual, word-shape and alignment models. It is completely unsupervised, with no manually labeled items, no external resources, only using parallel text that does not need to be easily alignable. The results are strong, with $F > 0.85$ for purely unsupervised named-entity recognition across languages, compared to just $F = 0.35$ on the same data for supervised cross-domain named-entity recognition within a language. A combination of unsupervised and supervised methods increases the accuracy to $F = 0.88$. We conclude that we have found a viable new strategy for unsupervised named-entity recognition across low-resource languages and for domain-adaptation within high-resource languages.

1 Introduction

At first pass, our approach sounds like it shouldn't work, as *'unsupervised'* tasks significantly underperform their supervised equivalents and for most cross-linguistic tasks *'unaligned'* will mean *'unusable'*. However, even among very loosely aligned multilingual text it is easy to see why named-entities are different: they are the least likely words/phrases to change form in translation. We can see this in the following example which shows the named-entities in both a Krèyol message and its English translation:

> Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.
>
> Sacre-Coeur Hospital which located in this village Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The example is taken from the parallel corpus of English and Haitian Krèyol text messages used in the *2010 Shared Task for the Workshop on Machine Translation* (Callison-Burch et al., 2011), which is the corpus used for evaluation in this paper.

The similarities in the named-entities across the translation are clear, as should be the intuition for how we can leverage these for named-entity extraction. Phrases with the least edit distance between the two languages, such as *'Lopital Sacre-Coeur'*, *'Milot'*, and *'Okap'*, can be treated as high-probability named-entity candidates, and then a model can be bootstrapped that exploits predictive features, such as word shape (e.g.: more frequent capitalization) and contextual cues such as the preceding *'vil'* in two cases above.

However, the problem of identifying entities in this way is non-trivial due to a number of complicating factors. The inexact translation repeats the non-entity *'hospital'* which limits machine-translation-style alignments and has an equal edit-distance with the entity *'Loptial'*. The entity *'Hospital'* and *'Lopital'* are not an exact match and are not perfectly aligned, changing position within the phrase. The

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 21–29,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

capitalization of entities is not always so consistent (here and in short-message communications more generally). A typographic error in the translation writes *'Okap'* as *'Oakp'*. *'Okap'* is itself slang for *'Cap-Haïtien'* and other messages translated this location across the different spellings (*'Cap-Haitien'*, *'Cap Haitien'*, *'Kap'*, *'Kapayisyen'*, etc.), which increases the edit distance. There are few resources for Haitian Krèyol such as gazatteers of place names (except at the Department/major Town/City level – at the time these messages were sent, *Google Maps* and *Open Street Map* listed only a handful of locations in Haiti, and such resources tend not to include slang terms). Finally, what was one sentence in the original message is split into two in the translation.

As Kay points out, most parallel texts *shouldn't* be alignable, as different contexts mean different translation strategies, most of which will not result in usable input for machine translation (Kay, 2006). This is true of the corpus used here – the translations were made for quick understanding by aid workers, explaining much of the above: it was clearer to break the translation into two sentences; it reduced ambiguity to repeat 'hospital' rather than leave it underspecified; the typo simply didn't matter. We confirmed the 'unalignability' of this corpus using the *GIZA++* aligner in the *Moses* toolkit (Koehn et al., 2007); by noting *Microsoft Research*'s work on the same data where they needed to carefully retranslate the messages for training (Lewis, 2010); and from correspondence with participants in the *2011 Workshop on Machine Translation* who reported the need for substantial preprocessing and mixed results.

We do not rule out the alignability of the corpus altogether – the system presented here could even be used to create better alignment models – noting only that it is rare that translations can be used straight-of-the-box, while in our case we *can* still make use of this data. Even with perfect alignment, the accuracy for named-entity extraction in Haitian Krèyol could only be as accurate as that for English, which in this case was $F = 0.336$ with a supervised model, so alignment is therefore only part of the problem.

For the same reasons, we are deliberately omitting another important aspect of cross-linguistic named-entity recognition: *transliteration*. Latin Script may be wide-spread, especially for low resource languages where it is the most common script for transcribing previously non-written languages, but some of the most widely spoken languages include those that use Arabic, Bengali, Cyrillic, Devanagari (Hindi) and Hanzi (Chinese) scripts, and the methods proposed here would be even richer if they could also identify named entities across scripts. A first pass on cross-script data looks like it *is* possible to apply our methods across scripts, especially because the seeds only need to be drawn from the most confident matches and across scripts there seem to be some named entities that are more easier to transliterate than others (which is not surprising, of course – most cross-linguistic tasks are heterogeneous in this way). However, with a few notable exceptions like Tao et al. (2006), transliteration is typically a supervised task. As with machine translation it is likely that the methods used here could aid transliteration, providing predictions that can be used within a final, supervised transliteration model (much like the semi-supervised model proposed later on).[1]

1.1 The limitations of edit-distance and supervised approaches

Despite the intuition that named-entities are less likely to change form across translations, it is clearly only a weak trend. Even if we assume oracle knowledge of entities in English (that is, imagining that we have perfect named-entity-recognition for English), by mapping the lowest edit-distance phrase in the parallel Krèyol message to each entity we can only identify an entity with about 61%, accuracy. *Without* oracle knowledge – training on an existing English NER corpora, tagging the English translations, and mapping via edit distance – identifies an entity with only around 15% accuracy. This is not particularly useful and we could probably achieve the same results with naive techniques like cross-linguistic gazetteers.

Edit distance and cross-linguistic supervised named-entity recognition are *not*, therefore, particularly useful as standalone strategies. However, we are able to use aspects of both in an unsupervised approach.

[1] On a more practical level, we also note that this year's shared task for the Named Entity Workshop is on transliteration. With the leading researchers in the field currently tackling the transliteration problem, it is likely that any methods we presented here would soon be outdated.

In this paper we focus on named-entity identification, only briefly touching on named-entity classification (distinguishing between types of entities), primarily because the named-entity identification component of our system is more novel and therefore deserves greater attention.

We use 3,000 messages in Haitian Krèyol and their English translations, with named-entities tagged in an evaluation set of 1,000 of the messages. To keep the task as unsupervised a possible, the system was designed and parameters were set without observing the actual tags.

1.2 Strategy and potential applications

Our approach is two-step for pairs of low resource languages, and three-step for pairs of languages where one has named-entity resources:

1. *Generate seeds by calculating the edit likelihood deviation*. For all cross-language pairs of messages, extract the cross-language word/phrase pairs with the highest edit likelihood, normalized for length. Calculate the intramessage deviation of this edit likelihood from the mean pair-wise likelihood from all candidate pairs within the message. Across all messages, generate seeds by selecting the word/phrase pairs with the highest and lowest intramessage edit likelihood deviation.
2. *Learn context, word-shape and alignment models*. Using the seeds from Step 1, learn models over the context, word-shape and alignment properties (but not edit distance). Apply the models to all candidate pairs. Because we have the candidate alignments between the languages, we can also jointly learn to identify named-entities by leveraging the context and word-shape features in the parallel text, in combination with the alignment predictions.
3. *Learn weighted models over the context, word-shape, alignment and supervised predictions (with high-resource languages only).* Using the seeds from Step 1 and predictions from Step 2, learn models over the broader features and supervised predictions from a model in the high-resource language, applying the models to all candidate pairs.

The results are very strong, with $F > 0.85$ for purely unsupervised named-entity recognition across languages. This is compared to just $F = 0.35$ for supervised approaches across domains within a language (MUC/CoNLL-trained English applied to the English translations of the messages).

The combined unsupervised/supervised methods increase the accuracy to $F = 0.88$. Inter-annotator agreement is around 0.95, so this may be close to the best possible result.

This leads us to conclude that cross-linguistic unsupervised named-entity recognition, even when not alignable via machine-translation methods, is a powerful, scalable technique for named-entity recognition in low resource languages.

The potential applications of are broad. There are some 5,000 languages in the connected world, most of which will have no resources *other* than loose translations, so there is great application potential. For high-resource languages, the results here indicate that the technique can be used to increase accuracy in cross-domain named-entity recognition, a consistent problem across even closely-related domains. For the specific corpus used there is also direct practical value – the messages include high volumes of time-critical requests for aid, citing locations that did not appear on any map in a language with few resources.

2 STEP 1: Establish Edit Likelihood Deviation

As we state in the introduction, we cannot simply tag in English and then find the least-edit distance word/phrase in the parallel Krèyol.

We evaluated several different edit distance functions, including the well-known Levenshtein and slightly more complex Jaro-Winkler measures. We also extended the Levenshtein measure by reducing the edit penalty for pairs of letters of phonetic relatedness, such as '*c*' and '*k*', following the subword modeling work of Munro and Manning on this corpus and previous subword modeling for short messages (Munro, 2011; Munro and Manning, 2010).[2]

[2] We also attempted a more sophisticated approach to learning weights for edits by extracting edit probabilities from the final model. This also made little improvement, but it could have simply been the result data-sparseness over only 3000 pairs of entities, so no strong conclusions can be drawn.

The more sophisticated edit distance functions gave more accurate predictions (which is unsurprising), but the advantages were lost in the following step when calculating the deviation from the norm, with all approaches producing more or less the same seeds. Rather than the String Similarity Estimate being the key factor, we conclude that our novel treatment of edit distance (calculating the local deviation) is the critical factor in generating seeds for the model.

All else being equal, then, we report results from the *simplest* approach to edit distance, normalizing Levenshtein's measure, $LEV()$ by length to a 0-1 scale. Candidate words/phrases were limited to a maximum of four words, delimited by space or punctuation, simply to cap the cost of the $LEV()$. Given a string S in message M, M_S and and its candidate pair $M'_{S'}$, and a length function $LEN()$, this gives us $SSE(M_S, M'_{S'}) =$

$$1 - \frac{(2(LEV(M_S, M'_{S'})) + 1}{LEN(M_S) + LEN(M'_{S'}) + 1}$$

The $+1$ smoothing is to avoid too much variation at smaller lengths, which is fairly common practice in subword models looking at morphological variation (Tchoukalov et al., 2010).

The String Similarity Estimate is a global measure that is not sensitive to the contexts of the given pairs. Suppose a sentence *wasn't* a translation, but simply a repetition, or that much of the translation was a direct (non-translated) quote of the original. Both occur in the data we used.

We propose, then, that the best candidate seeds for named-entities are those that display the highest likelihood relative to the other candidate pairs within the same pairs of messages. In other words, when there are two phrases with very little edit distance, but when there is very high cross-language edit distance between the contexts of the phrases. We define this as *Edit Likelihood Deviation*, $ELD()$.

There are many ways to calculating deviation. Again, to keep it as simple as possible we report results using the most well-known deviation metric, z-scores. Given average and standard deviation functions $AV()$ and $SD()$, gives $ELD(M_S, M'_{S'}) =$

$$\frac{(SSE(M_S, M'_{S'})) - AV(SSE(M_{0-n}, M'_{0-m}))}{SD(SSE(M_{0-n}, M'_{0-m}))}$$

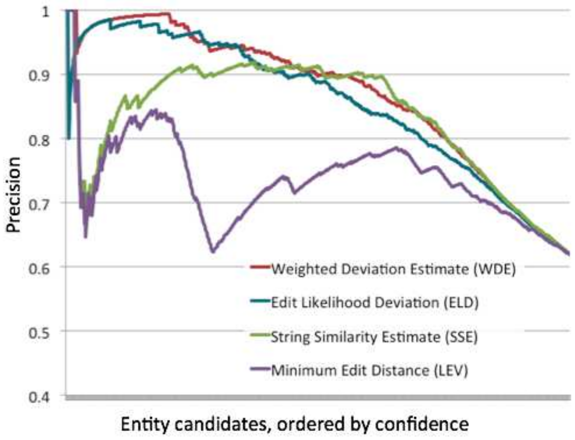

Figure 1: A comparison of the different approaches to generating seeds from edit distance. The comparison shows that local deviation, the novel method introduced in this paper, is the most successful. With about 10% of the most confident entity candidates by Edit Likelihood Deviation or Weighted Deviation Estimate, there is greater than 95% precision, giving a clean enough division of the data to seed a model.

At this point, we have the global string similarity of each candidate entity pair across languages, $SSE()$, and the local string similarity deviation of each candidate pair, $ELD()$.

A combination was also explored that combined the two, creating an equally weighted product of SSE and $ELD()$, *Weighted Deviation Estimate*, $WDE()$ (equation omitted for space). As Figure 1 shows, there is only a slight improvement from the combination of the two, showing that *Edit Likelihood Deviation*, the novel approach here, contributes the most to identifying candidate seeds.

We can calculate the first accuracies here by assuming that the best candidate in each message pair was an entity. All results also summarized at the end of the paper:

	Precision	**Recall**	**F-value**
Krèyol:	0.619	0.619	0.619
English:	0.633	0.633	0.633

The results are reasonably strong for methods that made few assumptions about the data and were not optimized, with errors in a little under half the predictions.

While the different equations are monotonically distributed *within* each pair of messages, the estimates *between* messages now take into account both local and global edit likelihoods, allowing us to rank the candidates by WDE and sample the most likely and least likely. Here, we simply took the top and bottom 5%.[3]

3 STEP 2: Learn joint alignment and word-shape models using the likelihood estimates as seeds.

Taking the seeds from Step 1, we can then treat them as training items in a linear model.

We used the Stanford Maximum Entropy Classifier. Model-choice is only important in that a discriminative learner is required. The 5% 'non-entity' pairs were still the highest String Similarity for their particular message/translation, but simply did not deviate greatly from the average within that message/translation. Therefore, we are explicitly targeting the border between entities and non-entities in the high String Similarity part of the vector space. This sampling strategy would not work for a generative learner.

For the same reason, though, we do *not* include raw edit distance or the String Similarity Estimate among the features. If we did, then the model will simply relearn and overfit this bias and give all the weight to edit distance.

We build the model on features that include context (the entity itself and surrounding words), word-shape features (capitalization, punctuation, segmentation, and numerical patterns), and alignment (absolute and relative character offsets between the candidates in the messages and translation). For word-shape features, we used a simple representation that converted all sequences of capitalized letters, lower-case letters, and non-letter characters into 'C', 'c' and 'n', respectively. Therefore, 'Port-au-Prince', 'Port au Prince' and 'Port.a.Prons' would all get the same word-shape feature, 'CcncnCc'. We allowed the model to also find character-ngrams over these shapes to capture features which would represent characteristics like 'is-capitalized', 'contains-internal-capital', and 'is-multiword-phrase'.

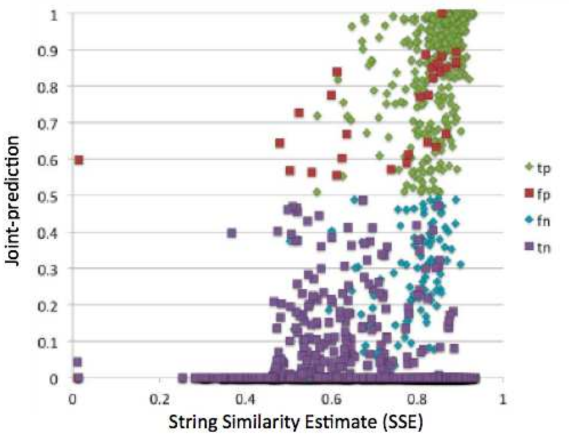

Figure 2: Comparing the predictions for the String Similarity for the same candidates, to the jointly-learned model. (Coding scheme: tp = true-positive, etc.) The distribution shows that while String Similarity correlates with named-entities, it is not a clean division. Note especially the mass of true-negatives in the bottom-right corner of the graph. These would be a relatively high volume of false-positives for String Similarity alone, but the model that bootstraps knowledge of context, word-shape and alignment has little trouble distinguishing them and correctly assigning them zero-probably of being an entity.

As a relatively small set of features, we also model the intersection of each of them. This allows the model to learn, for example, that words that are perfectly aligned, but are both all lower-case, are weighted 0.06 more likely as a non-entity. Despite the simplicity and low number of features, this is a fairly powerful concept to model.

As with all unsupervised methods that bootstrap predictions through seeded data, the success relies on a representative feature space to avoid learning only one part of the problem. The results are strong:

	Precision	**Recall**	**F-value**
Krèyol:	0.907	0.687	0.781
English:	0.932	0.766	0.840

There is a reasonably high precision-recall ratio which is typical of unsupervised learning that learns a model on seeded data, but the results are still

[3] There are clearly many more parameters and variants of equations that could be explored. As an unsupervised approach, it is by conscious choice that only the most well-known equations are used and tunable parameters are set at sensible defaults (like the equal weights here). This is to keep the experiments as cleanly 'unsupervised' as possible, and to demonstrate that the accurate results here are not simply a quirk of a particular equation, but a broadly applicable approach to generating seeds by local deviation estimates.

strong for both Krèyol and English, indicating that the seeding method in Step 1 did, in fact, produce candidates that occurred in broad range of contexts, overcoming one of the limits of gazetteer-based approaches.

Perhaps the most obvious extension is to jointly learn the models on both languages, using the candidate alignment models in combination with the contexts in both the original text and the translation:

	Precision	**Recall**	**F-value**
Krèyol:	0.904	0.794	0.846
English:	0.915	0.813	0.861

This improves the results for both, especially the Krèyol which can now take advantage of the more consistent capitalization and spelling in the English translations.

For many supervised learners, 0.846 would be a strong result. Here, we are able to get this in Hatian Krèyol using only unsupervised methods and a few thousand loosely translated sentences.

4 STEP 3: Learning weighted models over the context, word-shape, alignment and supervised predictions (with high-resource languages)

The natural extension to the supervised comparison is to combine the methods. We included the Stanford NER predictions in the features for the final model, allowing the bootstrapped model to arrive at the optimal weights to apply to the supervised predictions in the given context.

From the perspective of supervised NER, this can be thought of as leveraging unsupervised alignment models for domain-adaptation. The Stanford NER predictions were added as features in the final model, directly for the English phrases and across the candidate alignments for the Krèyol phrases.

Taken alone, the unsupervised strategies clearly improve the results, but for someone coming from a supervised learning background in NER (which will be most NER researchers) this should provide an intuition as to exactly how good. We cannot compare the Krèyol as there is no supervised NER corpus for Krèyol, and our labeled evaluation data is too small to train on. However, we can compare the English results to near state-of-the-art NER taggers.

We compared our system to the predictions made by the Stanford NER parser trained on MUC and CoNLL data (Sang, 2002; Sang and De Meulder, 2003):

	Precision	**Recall**	**F-value**
English:	0.915	0.206	0.336

The low cross-domain result is expected, but 0.336 for supervised cross-domain predictions within a language is *much* less than 0.861 for unsupervised cross-language predictions. This clearly shows that the methods and evaluation used here really do demonstrate a new strategy for NER. It also shows that domain-specificity might be even be more important than language-specificity when we can bootstrap our knowledge of context.[4]

Combining the two approaches, we get the most accurate results:

	Precision	**Recall**	**F-value**
Krèyol:	0.838	0.902	0.869
English:	0.846	0.916	0.880

Even though English is a high-resource language, this is still a very good result for cross-domain adaptation, with $F > 0.5$ improvement over the supervised model alone. It is clear that this strategy could be used for domain adaptation more broadly wherever loose translations exists.

While not as big a gain in accuracy as the previous steps, the $F > 0.02$ gain is still significant. Although untested here, it is easy to imagine that with a small amount of labeled data or improved gazetteers the supervised approach should further. About 10% of the error can be attributed to capitalization, too, which is a slight bias against the MUC/CoNLL trained data where the capitalization of named entities was consistent. A realistic deployment approach would be to create an initial model using the unsupervised methods described in this paper and then to further bootstrap the accuracy through supervised labeling. This particular approach to semi-supervised learning is outside the scope of this paper.

[4] For the edge cases and entity boundary errors, we always gave the benefit of the doubt to the Stanford NER tagger.

4.1 Distinguishing Types of Entity

NER often distinguishes types of Entities (eg: People, Locations, Organizations); a frequent subtask sometimes called *named-entity discrimination* or *named-entity classification*. We discuss this briefly.

By seeding the data with the Stanford NER predictions for 'Person', 'Location', and 'Organization' and learning a three-way distinction within the entities, we saw that it wasn't a difficult problem for this particular corpus. The main potential complication was between organizations and locations (especially for radio stations) but there were relatively few organizations in the data so the micro-fvalue would change very little. No doubt, in other texts the location/organization division would compose a bigger part of the problem. These observations about distinguishing NERs are consistent with the known problems in NER more broadly. The Stanford NER only made predictions for 114 of the entities that were confidently mapped to their Krèyol counterparts in Step 1:

	Precision	**Recall**	**F-value**
English:	0.512	0.782	0.619

To exploit *any* signal here, let alone a respectable $F = 0.619$ is a good result, but clearly more improvements are possible.

5 Analysis

The results presented in the paper are summarized in Table 1. Taken together, they make it clear that this is a very promising new method for named-entity recognition in low resources languages, and for domain-adaptation in high-resource languages.

Analysis of the consistent errors shows several clear patterns. Products like *'aquatab'* were a common false positive, although a product could be a named-entity in certain coding schemas. Dates, figures and currency (*'250gd'*) were also frequent false positives, but would be reasonably easy to filter as they follow predictable patterns.

Some cognates and borrowings also made it through as false-positives: *'antibiotics'*/*'antibiotik'*, *'drinking water'*/*'drinking water'*, *'medicine'*/*'medicament'*, *'vitamin c'*/*'vitamine c'*, *'cyber'*/*'cyber'*, *'radio'*/*'radyo'*, although *'cyber cafe'* almost always referred to a specific location and *'radio'* was often part of an organization name, *'radio enspirasyon'*.

The false-negatives were almost all very low-frequency words or high-frequency words that were more commonly used as non-entities. This is consistent with named-entity recognition more broadly.

6 Background and Related Work

We were surprised that no one had previously reported looked at leveraging cross-linguistic named-entity recognition in this way. Perhaps previous researchers had found (like us) that edit distance alone was not particularly useful in cross-linguistic named-entity recognition, and therefore not pursued it. While the approach is novel, the general observation that named-entities change form less than other words cross-linguistically is one of the oldest in language studies. Shakespeare's 'River Avon' simply means 'River River', as 'Avon' is, literally, 'River' in the pre-English Celtic language of the region.

For parallel short-message corpora, named-entity recognition is completely unresearched, but there is growing work in classification (Munro and Manning, 2010; Munro, 2011) and translation (Lewis, 2010), the latter two using the same corpus as here.

Past *'Multilingual Named-Entity Recognition'* systems meant training the same supervised system on different languages, which was the focus of the past CoNLL tasks. While the goal of these systems was the same as ours – broad cross-linguistic coverage for named-entity recognition – this is *not* the same 'cross-linguistic' as the one employed here.

More closely related to our work, Steinberger and Pouliquen have found cross-linguistic named-entity recognition to be possible by aligning texts at the granularity of news stories (Steinberger and Pouliquen, 2007), but using a supervised approach for the first pass and focusing on transliteration. In other related work, the 2007 NIST *REFLEX* evaluation (Song and Strassel, 2008), tasked participants with using alignment models to map named-entities between English, Arabic, and Chinese data. They found that relying on alignment models alone was very poor, even among these high-resource languages, although it was a relatively small corpus (about 1,000 aligned entities). The focus was more on transliteration – an important aspect of translation that we simply aren't addressing here.

Most earlier work used a tagger in one language in combination with machine translation-style alignments models. Among these, Huang et al. is the most closely related to our work as they are translating rare named-entities, and are therefore in a similar low-resource context (Huang et al., 2004). As with the *NIST* project, most work building on Huang et al. has been in transliteration.

Although not cross-linguistic, Piskorski et al.'s work on NER for inflectional languages (2009) also relied on the similarities in edit distance between the *intra*-language variation of names.

In gazetteer-related work, Wang et al. and others since, have looked at edit distance within a language, modeling the distance between observed words and lists of entities (Wang et al., 2009). Similarly, there is a cluster of slightly older work on unsupervised entity detection, also within one language (Pedersen et al., 2006; Nadeau et al., 2006), but all relying on web-scale quantities of unlabeled data.

While the implementation is not related, it is also worth highlighting Lin et al.'s very recent work on unsupervised language-independent name translation the mines data from Wikipedia 'infoboxes', (Lin et al., 2011) however the infoboxes give a fairly and highly structured resource, that might be considered more supervised than not.

In alignment work, the foundational work is Yarowsky et al.'s induction of projections across aligned corpora (Yarowsky et al., 2001), most successfully adapted to cross-linguistic syntactic parsing (Hwa et al., 2005). The machine translation systems used named-entity recognition are too many to list here, but as we say, the system we present could aid translation considerably, especially in the context of low resources languages and humanitarian contexts, a recent focus in the field (Callison-Burch et al., 2011; Lewis et al., 2011).

7 Conclusions

We have presented a promising a new strategy for named-entity recognition from unaligned parallel corpora, finding that unsupervised named-entity recognition across languages can be bootstrapped from calculating the local edit distance deviation between candidate entities. Purely unsupervised approaches are able to identify named entities with $F = 0.846$ accuracy for Krèyol and $F = 0.861$ for English, leveraging the candidate alignments for improved accuracy in both cases. Combined with supervised learning, the accuracy rises to $F = 0.869$ and $F = 0.880$ respectively, which is approaching the level of accuracy achieved by in-domain supervised systems. It is rare for unsupervised systems to be competitive with supervised approaches as accuracy is usually lost for coverage, but here it looks like the method can be effective for both.

Unsupervised	**Precision**	**Recall**	**F-value**
Edit likelihood deviation			
Krèyol:	0.619	0.619	0.619
English:	0.633	0.633	0.633
Language-specific models			
Krèyol:	0.907	0.687	0.781
English:	0.932	0.766	0.840
Jointly-learned models			
Krèyol:	0.904	0.794	**0.846**
English:	0.915	0.813	**0.861**
Supervised			
English:	0.915	0.206	0.336
Semi-supervised			
Identification			
Krèyol:	0.838	0.902	**0.869**
English:	0.846	0.916	**0.880**
Classification (micro-F)			
English:	0.512	0.782	0.619

Table 1: A summary of the results presented in this paper showing promising new methods for unsupervised and semi-supervised named-entity recognition.

There is the potential to apply this system to a large number of natural language processing problems, and to extend the system in a number of directions. Each of the three steps has parameters that could be optimized, especially in combination with supervised approaches. The linguistic nature of the language pairs might also influence the effectiveness. The results here are therefore the first presentation of a new strategy – one that will hopefully lead to more research in extracting rich information from a diverse range of low-resource languages.

References

C.. Callison-Burch, P. Koehn, C. Monz, and Zaidan. O. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

F. Huang, S. Vogel, and A. Waibel. 2004. Improving named entity translation combining phonetic and semantic similarities. In *Proc. of HLT-NAACL*, pages 281–288.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

M. Kay. 2006. Translation, Meaning and Reference. *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, page 3.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

W. Lewis, R. Munro, and S. Vogel. 2011. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

W. Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual Conference of the European Association for Machine Translation*.

W.P. Lin, M. Snover, and H. Ji. 2011. Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes. *Proceedings of the EMNLP Workshop on Unsupervised Learning in NLP*, page 43.

R. Munro and C.D. Manning. 2010. Subword variation in text message classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*.

R. Munro. 2011. Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*.

D. Nadeau, P. Turney, and S. Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence*, pages 266–277.

T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio. 2006. An unsupervised language independent method of name discrimination using second order co-occurrence features. *Computational Linguistics and Intelligent Text Processing*, pages 208–222.

J. Piskorski, K. Wieloch, and M. Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299.

E.F Tjong Kim Sang and F. 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-Volume 4*, pages 142–147. Association for Computational Linguistics.

E.F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, proceedings of the 6th conference on natural language learning. *August*, 31:1–4.

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

R. Steinberger and B. Pouliquen. 2007. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162.

T. Tao, S.Y. Yoon, A. Fister, R. Sproat, and C.X. Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 250–257. Association for Computational Linguistics.

T. Tchoukalov, C. Monson, and B. Roark. 2010. Morphological analysis by multiple sequence alignment. *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 666–673.

W. Wang, C. Xiao, X. Lin, and C. Zhang. 2009. Efficient approximate entity extraction with edit distance constraints. In *Proceedings of the 35th SIGMOD international conference on Management of data*, pages 759–770. ACM.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.