A simple feature-copying approach for long-distance dependencies

Marc Vilain, Jonathan Huggins, and Ben Wellner
The MITRE Corporation
202 Burlington Rd
Bedford, MA 01730 (USA)
{mbv,jhuggins,wellner}@mitre.org

Abstract

This paper is concerned with statistical methods for treating long-distance dependencies. We focus in particular on a case of substantial recent interest: that of long-distance dependency effects in entity extraction. We introduce a new approach to capturing these effects through a simple feature copying preprocess, and demonstrate substantial performance gains on several entity extraction tasks.

1 Long-distance dependencies

The linguistic phenomena known as long-distance dependencies have a long history in computational linguistics. Originally arising in phrase-structure grammar, the term aptly describes phenomena that are not strictly grammatical, and has thus gained currency in other endeavors, including that of concern to us here: entity extraction. The common thread, however, is simply that the treatment of a linguistic constituent α might be influenced by the treatment of a non-local constituent β.

In phrase-structure grammar, dependencies arise between matrix phrases and the gapped phrases that they dominate, as in "the cake that I hope you'll serve ε". The idea that these are *long-distance* dependencies arises from the fact that the separation between linked constituents can be arbitrarily increased while their dependency continues to hold (as in "the cake that I hope you'll ask Fred to tell Joan to beg Maryanne to serve ε").

With entity extraction, long-distance dependencies typically occur between mentions of the same entity. Consider, for example, the italicized references to Thomas White in this newswire excerpt:

> Bank of America on Friday named *Thomas White* head of global markets. *White* has been global head of credit products.

The fact that the first of these mentions is easily understood as person-denoting has substantial bearing on interpreting the second mention as person-denoting as well. But while local evidence for personhood is abundant for the first instance (*e.g.*, the given name "Thomas" or the verb "named"), the evidence local to the second instance is weak, and it is highly unlikely that a learning procedure would on its own acquire the relevant 5-gram context (α has been $\beta_{JJ}\ \gamma_{title}$). The dependency between these instances of *White* is thus a significant factor in interpreting both as names.

It is well known that capturing this kind of dependency can dramatically improve the performance of entity extraction systems. In this paper, we pursue a very simple method that enables statistical models to exploit these long-distance dependencies for entity extraction. The method obtains comparable or better results than those achieved by more elaborate techniques, and while we focus here on the specific case of entity extraction, we believe that the method is simple and reliable enough to apply generally to other long-distance phenomena.

2 Approaches to name dependencies

The problem of capturing long-distance dependencies between names has a traditional heuristic solution. This method, which goes back to systems participating in the original MUC-6 evaluation (Sundheim, 1995), is based on a found names list. The method requires two passes through the input. A first pass captures named entities based on local

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 192–200,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

evidence, and enters these names into a found names registry. A second pass identifies candidate entities that were missed by the first pass, and compares them to entries in the registry. Where there is string overlap between the candidate and a previously found name, the entity type assigned to the existing entry is copied to the candidate.

Overall, this is an effective strategy, and we used it ourselves in a rule-based name tagger from the MUC-6 era (Vilain and Day, 1996). The strategy's Achilles heel, however, is what happens when erroneous entries are added to the found names list. These can get copied willy-nilly, thereby drastically increasing the scope of what may originally have started as a single local error. Clearly, the approach is begging to be given a firmer evidence-weighing foundation.

2.1 A statistical hybrid

An early such attempt at reformulating the approach is due to Minkheev *et al* (1999). As with previous approaches, Mikheev and his colleagues use a rule-based first pass to populate a found-names list. The second pass, however, is based on a maximum entropy classifier that labels non-first-passed candidates based on evidence accrued from matching entries on the found-names list. The statistical nature of the decision eliminates some of the failure modes of the heuristic found-names strategy, and in particular, prevents the copying of single errors committed in the first pass. The major weakness of the approach, however, is the heuristic first pass. Minkheev *et al* note that their method is most effective with a high-precision found-names list, implemented as a tightly controlled (but incomplete) rule-based first pass.

2.2 Fully-statistical models

Several more recent efforts have attempted to remove the need for a heuristic first-pass tagger, and have thus cast the problem as one-pass statistical models (Bunescu and Mooney, 2004; Sutton and McCallum, 2004; Finkel *et al*, 2005). While the technical details differ, all three methods approach the problem through conditional random fields (CRFs). In order to capture the long-distance dependencies between name instances, these approaches extend the linear-chain sequence models that are typically used for extracting entities with a CRF (Sha and Pereira, 2003). The resulting models consist of sentence-length sequences interlinked on those words that might potentially have long-distance interactions. Because of the graph-like nature of these models, the simplifying assumptions of linear-chain CRFs no longer hold. Since complete parameter estimation is intractable under these conditions, these three approaches introduce approximate methods for parameter estimation or decoding (Perceptron training for the first, loopy belief propagation for the first two, Gibbs sampling and simulated annealing for the third).

Krishnan and Manning (2006) provide a lucid critique of these extended models and of their computational ramifications. In a nutshell, their critique centers on the complexity of constructing the linked graphs (which they deemed high), the stability of Perceptron training (potentially unstable), and the run-time cost of simulated annealing (undesirably high). Since these undesirable properties are directly due to the treatment of long-distance dependencies through graphical models, it is natural to ask whether graphical models are actually required to capture these dependencies.

2.3 Avoiding non-sequential dependencies

In point of fact, Krishnan and Manning (2006) present an alternative to these graph-based methods. In particular, they break the explicit links that mutually condition non-adjacent lexemes, and instead rely on separate passes in a way that is reminiscent of earlier methods. A first-pass CRF is used to identify entities based solely on local information. The entity labels assigned by this first CRF are summarized in terms of lexeme-by-lexeme majority counts; these counts are then passed to a second CRF in the form of lexical features.

Consider, for example, a financial news source, where we would expect that a term like "Bank" might be assigned a preponderance of ORG labels by the first-pass CRF. This would be signaled to the second-pass CRF through a *token majority* feature that would take on the value *ORG* for all instances of the lexeme "Bank". This effectively aggregates local first-pass labeling decisions that apply to this lexeme, and makes the second-pass CRF sensitive to these first-pass decisions. Further refinements capture cases where a lexeme's label diverges from the token majority, for example: "Left Bank," where "Bank" will be assigned a LOC-valued *entity majority* feature whenever it appears in that particular word sequence. By capturing long-distance dependencies through lexical features, Krishnan and Manning avoid the need for graphical models, thus regaining tractability.

How well does this work? Returning to our earlier example, the idea behind these majority count features is that a term like "White" might be assigned the PER label by the first CRF when it appears in the context "Thomas White." Say, for the sake of argument, that sufficiently many instances of "White" are labeled PER by the first pass to sum to a majority. The second-stage CRF might then be expected to exploit the majority count features for "White" to PER-label any instances of White that were left unlabeled in the first pass (or that were given erroneous first-pass labels).

The method would be expected to fail, however, in cases where the first pass yields a majority of erroneous labels. Krishnan and Manning suggest that this is a fairly unlikely scenario, and demonstrate that their approach effectively captures long-distance name dependencies for the CoNLL English name-tagging task. They measured a best-in-class error reduction of 13.3% between their two-pass method and a single-stage CRF equipped with comparable features.

3 A contradictory data set

Just how unlikely, however, is the majority-error scenario that Krishnan and Manning discount? As it turns out, we encountered precisely this scenario while working with a corpus that is closely related to the CoNLL data used by Krishnan and Manning.

The corpus in question was drawn from the online edition of Reuters business news. The articles cover a range of business topics: mergers and acquisitions (M+A), stock valuations, management change, and so forth. This corpus is highly pertinent to this discussion, as the CoNLL English data are also Reuters news stories, drawn from the general news distribution. Our business data thus represent a natural branch of the overall CoNLL data.

A characteristic of these Reuters business stories that distinguishes them from general news is the prevalence of organization names, in particular company names. In these data, instances of company names significantly outnumber the next-most-common entities (money, dates, and the like). Even state-of-the-art CRFs trained on these data therefore err on the side of generating companies, meaning that in the absence of countermanding evidence (such as the presence of a person's given name), an entity will tend to be labeled ORG by default. Our earlier "Thomas White" example is a case in point: where the full name would typically be labeled PER, last-name-only instances ("White") might go unlabeled or be marked ORGs.

Label accuracy	count	% test cases
Trivially correct (present in both test and training)	6	—
Majority correct, test only	13	45%
Majority incorrect, test only	11	38%
Equivocal, test only	5	17%

Table 1: effectiveness of majority counts as predictors of entity type, Reuters business news sample

Table 1, above, shows a qualitative analysis of this phenomenon for PER entities in our M+A test set. The table considers person-denoting entities with three or more instances in the test set (n=35), and summarizes the majority accuracy of the labels assigned to them by a feature-rich 1-pass CRF. Of these thirty-five cases, we eliminate from consideration six trivial test cases that are present unambiguously in the training data (*e.g.*, "Carl Icahn"), since the CRF will effectively memorizes these cases during training. Of the remaining twenty-nine non-trivial cases, not quite half of them (45%) were accurately labeled by the CRF for the majority of their instances. A larger number of entities either received an incorrect majority label (38%) or were equivocally labeled, receiving an equal number of correct and incorrect tags (17%).

For this data set then, majority count features are poor models of the long-distance dependencies between person names, as they are just about as likely to predict the wrong label as the correct one.

4 A feature-copying alternative

A further analysis of our business news test sample revealed an intriguing fact. While in the absence of compelling evidence, the CRF might label a mention of a person entity as an org (or leave it unlabeled), for those mentions where compelling evidence existed, the CRF generally got it right. By compelling evidence, we mean such linguistic cues as the presence of a given name, contextual proximity to agentive verbs (*e.g.* "said"), and so forth.

This suggests an alternative approach to capturing these kinds of long-distance dependencies between names. In contrast to previous approaches, what is needed is not so much a way of coordinating non-local *decisions* about an entity's label, as a way of coordinating non-local *evidence* pertinent to the labeling decision. That is, instead of conditioning the labeling decision of a lexeme on the labeling decisions for that lexeme elsewhere in the corpus, we ought to condition the decision on the key evidence supporting those decisions.

4.1 Displaced features

Our approach operates by identifying those features of a CRF that are most predictive over a corpus. Each of those features is then duplicated: for a given token α, one version of the feature applies directly to α, while the other version applies to all other instances where α's word form appears in the current document. In particular, what we duplicate is the indicator function for a feature. The local version of an indicator Φ signals true if it applies locally to α, while the displaced version Φ_d signals true if it applies to *any* token α' that is an instance of the same word from as α.

To make this concrete, consider our opening example, now indexed with word positions:

Thomas$_7$ *White*$_8$ … *White*$_{13}$ has$_{14}$ been$_{15}$ …

Say that Φ is a feature indicator that is true of a token α_i just in case the token to its left, α_{i-1}, is a given name. In this instance, $\Phi(\textit{White}_8)$ is true and $\Phi(\textit{White}_{13})$ is false. Then Φ_d, the displaced version of Φ, will be true of α_i just in case there is some token α_j with the same word form such that $\Phi(\alpha_j)$ is true. In this instance $\Phi_d(\textit{White}_8)$ and $\Phi_d(\textit{White}_{13})$ are both true by virtue of Φ being true of *White*$_8$.

This feature displacement scheme introduces non-local evidence into labeling decisions, effectively capturing the long-distance dependencies exhibited by name-tagging tasks. The method differs from previous approaches in that the models are not made conditional on non-local decisions (as in the case of graphical models), nor are they made conditional on aggregated first-pass decisions (as in Krishnan & Manning), but rather are made conditional on non-local evidence (displaced features).

4.2 Identifying features to displace

Because a typical entity extraction model can use tens or hundreds of thousands of features, it is not practical to displace every one of them. Though technically this only doubles the number of features under consideration, the lexical indexing rapidly gets out of hand. In addition, training and run times increase and, in our experience, a risk of over-fitting emerges. In point of fact, however, capturing long-distance name dependencies does not require us to replicate every last bit of feature-borne evidence. Instead, we only need to displace the evidence that is most reliably predictive.

To select predictive features to displace, we've had most success with a method based on information gain. Specifically, we use a one-time pre-process that measures feature gain relative to a corpus. The pre-process considers the same complement of feature schemas as are used by the actual CRF, and grounds the schemas on a training corpus to instantiate free lexical and P-O-S parameters. Gain for the instantiated features is measured through K-L divergence, and the *n* features with highest gain are then selected for displacement (with *n* typically ranging from 1,000 to 10,000).

As in (Schneider, 2004), gain for a given feature Φ, is found through a variant of the familiar Kullback-Leibler divergence formula,

$$D_{KL}(P \parallel Q) = \sum_i p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}$$

For our purposes, the x_i are the non-null entity labels defined for the training set (PER, ORG, *etc.*), *P* is the probability distribution of the labels over the training set, *Q* is the distribution of the labels over tokens for which Φ applies, and *p* and *q* are their respective smoothed probability estimates (Laplace smoothing). Note in particular that this formulation excludes the null label ("not an entity"). This effectively means that K-L divergence is giving us a measure of the degree to which a feature predicts one or more non-null entity labels. Because the null label is generally the dominant label in named-entity tasks, including the null label in the calculation of K-L divergence tends to overwhelm the statistics, and leads to the selection of uninformative features that predict non-entities.

Figure 1 demonstrates the effectiveness of this feature selection method, along with sensitivity to the threshold parameter. The figure charts F-score on a Reuters business news task (M+A) as a function of the number of displaced features. From a baseline of F=89.3, performance improves rapidly with the addition of displaced features to the CRF model, reaching a maximum of F=91.4 with the

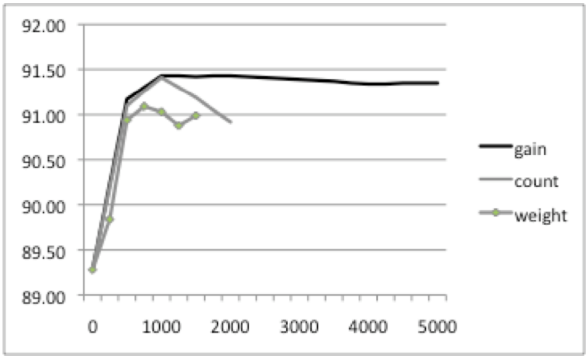

Figure 1: F score on the Reuters M+A task, as a function of number of displaced features

addition of 1,000 displaced features. Performance then fluctuates asymptotically around this level.

The chart also shows comparable growth curves for two alternative feature selection methods. The feature count method is similar to feature gain, but instead of ranking features with K-L divergence, it ranks them according to the number of times they match against the corpus. Feature weight does not use a schema-grounding first pass to generate candidate features, but trains a CRF model on the corpus, and then ranks features according to the weight assigned to them in the model. In preliminary experiments, neither of these methods yielded as high-performing a set of displaced features as feature gain. Additionally their growth curves exhibit sensitivity to parameter setting, which suggests a risk of over-fitting. For these reasons, we did not pursue these approaches further.

Note finally that the feature schemas we consider for displacement only encode local evidence (see Table 2 below). In particular, they do not encode the assigned label of a word form, as this would effectively introduce the kind of graphical conditional dependencies that lie outside the scope of linear-chain CRF methods.

4.3 Training and decoding

Aside from two pre-processing steps, training or decoding a CRF with displaced features is no different from training or decoding one with only conventional features. As to the pre-processing steps, the first applies to the corpus overall, as we must initially select a collection of locally predictive features to displace. The second step applies on a per-document basis and consists of the creation of the inverted lexical indices that are used to trigger indicator functions for displaced features.

While these additional steps complicate training and decoding somewhat, they have little effect on actual decoding run times. Most importantly, they retain the linear-chain properties of the CRF, and therefore do not require the graphical modeling and involved parameter estimation called for by most previous approaches. In addition, the training logistics are of a lesser magnitude than those required by Krishnan and Manning's approach, since training their second-stage model first requires round-robin training of one-fold-left-out classifiers that estimate first-stage majority counts.

5 Experimental design

To evaluate the effectiveness of feature copying with long-distance dependencies, we undertook a number of information extraction experiments. We focused on the traditional name-tagging task, relying on both current and archival data sets. For each data set, we trained entity-extraction models that corresponded to three different strategies for capturing long-distance dependencies.

- Baseline model: a feature-rich CRF trained with only local features and no long-distance dependency features;
- Feature-copying model: a CRF trained with the same local features, along with displaced versions of high-gain features;
- Majority model: a re-implementation of the Krishnan and Manning strategy, using the same feature set as the baseline CRF as well as their majority count features.

We used held-out development test sets to tune the selection of displaced features, in particular, the number of features to displace.

5.1 CRF configurations

We used the Carafe open-source implementation of sequence-based conditional random fields.[1] Carafe has achieved competitive results for standard sequence modeling tasks (Wellner & Vilain, 2006, Wellner *et al*, 2007), and allows for flexible feature design. Carafe provides several learning methods, including a fast gradient descent method using periodic step-size adjustment (Huang *et al*, 2007). Preliminary trials, however, produced better results

[1] http://sourceforge.net/projects/carafe

lexical unigrams	$w_{-2} \ldots w_{+2}$
lexical bigrams	$w_{-2},w_{-1} \ldots w_{+1},w_{+2}$
P-O-S unigrams	$p_{-2} \ldots p_{+2}$
P-O-S bigrams	$p_{-2},p_{-1} \ldots p_{+1},p_{+2}$
substrings	.*s or s.* $\|s\|\leq 4$
linguistic word lists	gazetteers, date atoms, …
regular expressions	caps., digits, …
“corp.” nearby	also “ltd.” …

Table 2: Baseline features; w_i and p_i respectively denote lexeme and P-O-S in relative position i.

Corpus	Language	NU	TM	MI	Topics
MUC-6	English	✓	✓		mostly politics
MUC-7	English	✓	✓r		mostly politics
MNET	Spanish	✓	✓r		mostly politics
Reuters	English	✓	✓		business
CoNLL	English			✓	all news

Table 3: Data set characteristics. All include persons, organizations, and locations; some have numeric forms (NU), dates and times (TM) where r indicates relative dates, or misc (MI).

with conditional log-likelihood learning (L-BFGS optimization). We used this latter method here, L2-regularized by a spherical Gaussian prior with variance set to 10.0 (based on preliminary trials).

Our baseline CRF was given a feature set that has proven its mettle in the literature (see Table 2). Along with contextual n-grams and the like, these features capture linguistic regularities through membership in vocabulary lists, *e.g.*, first names, major geographical names, honorifics, *etc.* They also include hand-engineered lists from our legacy rule-based tagger, *e.g.*, head word lists for organization names, lists of agentive verbs that reliably apply to persons, date atoms, and more. For part-of-speech features, we either accepted the parts of speech provided with a data set, or generated them with our implementation of Brill's method (Brill, 1994). For the majority count features, we used document and corpus versions the *token* and *entity* features described by Krishnan and Manning, but did not re-implement their *super-entity* feature.

5.2 Experimental data

We evaluated our approach on five different data sets: our current corpus of Web-harvested Reuters business news, as well as four archival data sets that have been reported on by other researchers. The business news data consist of a training corpus of mergers and acquisition stories (M+A), development and evaluation test sets for M+A and test sets for three additional topics: hot stocks (HS), new initiatives (NI), and general business news (BN). Table 3 provides an overview of our data sets and of some salient distinctions between them.

All five extraction tasks require the reporting of three core entity types: persons, organizations, and locations; additional required types are noted in the table. The reporting guidelines for the first four tasks are closely related: Reuters business and MUC-6 were annotated to the same original MUC-6 standard, while MUC-7 and MNET extend the MUC-6 standard slightly. The CoNLL standard alone calls for a catch-all (and troublesome) MISC entity.

5.3 Scoring metrics

Previous results on these data sets have been reported using one of two scoring methods: strict match (CoNLL) or match with partial credit, as calculated by the MUC scorer (MUC-6, MUC-7, and MNET). To enable comparisons to previously published work, we report our results with the metric appropriate to each data set (we use the MUC scorer for Reuters). These scoring distinctions are pertinent only to comparisons of absolute performance. In this paper, the interest is with relative comparisons across approaches to long-distance dependencies, for which the scorers are kept constant.

6 Experimental results

Table 4 summarizes our experimental results for the seven test sets annotated to the MUC-6 standard or its close variants (we will consider the CoNLL task separately). Along with F scores for our baseline CRF, the table presents F scores and baseline-relative error reduction (Δ_E) for two approaches to long-distance name dependencies: feature displacement (disp) and the Krishnan and Manning strategy (K+M). We were pleased to see that feature displacement proved effective for all of the extraction tasks. As the table shows, the addition of displaced features consistently reduced the residual error term left by the baseline CRF trained only with local features. For the English-language corpora, the error reduction ranged from a low of 11 % for the Reuters NI task to a high of 39% for the MUC-6 task. The error reduction for the Spanish-language MNET task was lowest of all, at 8.9%.

For all the English tasks, we consistently achieved better results with feature displacement

	MUC-6		MUC-7		MNET		Reuters M+A		Reuters BN		Reuters HS		Reuters NI	
	F	Δ_E	F	Δ_E	F	Δ_E	F	Δ_E	F	Δ_E	F	Δ_E	F	Δ_E
baseline	88.2	—	84.0	—	88.9	—	89.3	—	89.5	—	85.4	—	88.8	—
disp.	**92.8**	**39%**	**86.2**	**14%**	89.9	8.9%	**91.4**	**20%**	**91.8**	**22%**	**87.3**	**13%**	**90.1**	**11%**
K+M	91.5	28%	85.2	7.4%	—	—	90.4	11%	91.0	14%	86.3	6.2%	89.2	2.8%

Table 4: Performance on seven test sets annotated to variants of the MUC-6 standard (MUC scorer).

than with our version of Krishnan and Manning's approach (we were not able to obtain Spanish K+M results by publication time). In each case, displacement produced a greater reduction in baseline error than did majority counts. Furthermore, because both approaches start from the same baseline CRF, the resulting raw performance was consequently also higher for displacement. Note in particular the Reuters M+A test set: these are the data for which Table 1 suggests that majority counts would be poor predictors of long-distance effects. This prediction is in fact borne out by our results.

6.1 Effects of linguistic engineering

We were interested to note that the feature displacement method achieved both highest performance and highest error reduction for the MUC-6 corpus (F=92.8, Δ_E=39.3%) and for two of the Reuters test sets: M+A (F=91.4, Δ_E=20.0%) and BN (F=91.8, Δ_E=21.6%). The MUC-6 F-score, in particular, is comparable to those of hand-built MUC-era systems; in fact, it *exceeds* the score of our own hand-built MUC-6 system (Aberdeen *et al*, 1995).

What is apparently happening is that these three data sets are well matched to a group of linguistically inspired lexical features with which we trained our baseline CRF. In particular, our baseline features include gazetteers and word lists hand-selected for identifying entities based on local context: first names, agentive verbs, date atoms, *etc*. This played out in two significant ways. First, these linguistic features tended to elevate baseline performance (see Table 4). Second, these same features also proved effective when displaced, as demonstrated by the substantial error reduction with displacement. Feature displacement thus further rewards sound feature engineering.

6.2 Other MUC-related results

The MUC-7 and Reuters hot stocks data (HS) provide informative contrasts. For these data, feature displacement provided error reduction of Δ_E=13.9% and 13.4% respectively, which is less than for the top three data sets. It is interesting to note that in both cases, the baseline score is also lower, suggesting again that the performance of feature copying follows the performance of baseline tagging. In the case of Reuters HS, the evaluation data contained many out-of-training references to stock indices, which depressed baseline scores. Similar development-to-evaluation divergences have also been noted with the MUC-7 corpus.

6.3 The CoNLL task

Our results for the CoNLL task, reported in Table 5 below, provide a different point of contrast. The middle two rows of the table present the same experimental configurations as have been discussed so far. For this data set, we note that feature displacement does not perform as well as our reimplementation of Krishnan and Manning's strategy in terms of both absolute score and error reduction. Likewise, published results for other approaches mostly outperform displacement (see the first three rows in Table 5).

One possible explanation lies with the linguistic features with which we approached CoNLL: these are the same ones we originally developed for MUC-6. As noted earlier the CoNLL standard diverges in several ways from MUC-6. In particular, CoNLL calls for a MISC entity that covers a range of name-like entities, *e.g.*, events. MISC also, however, captures names that are trapped by tokenization ("London-based"), as well as some MUC organizations (sports leagues). This suggests that adapting our features to the CONLL task might help.

	base F	LDD F	Δ_E
Bunescu + Mooney 2004	80.09	82.30	11.1%
Finkel et al 2005	85.51	86.86	9.3%
Krishnan + Manning 2006	85.29	87.34	13.3%
K+M (re-impl, MUC feats.)	84.3	86.0	10.7%
displacement (MUC feats.)	84.3	85.8	9.6%
displ. (CoNLL feats.)	85.24	86.55	8.9%
displ. (CoNLL feats. + DS)	86.57	87.39	6.1%

Table 5: Performance on the CoNLL task; LDD designates use of long-distance dependency method.

The final two rows in Table 5 present attempts to tune our features to CoNLL. This includes some features (the "CoNLL feats" in Table 5) indicating story topic, all-caps headline contexts, presence in a sporting result table, and similar idiosyncrasies. In addition, we also used features based on distributional similarity word lists (DS in the table) provided with the Stanford NER package.[2]

While these feature engineering efforts proved effective, what we found surprised us. As Table 5 shows, the CoNLL features do substantially raise baseline performance, with the full set of new features producing a baseline (F=86.6) that outperforms previously published baselines by over a point of F score. In keeping with our observations for the MUC-annotated text, we would then have expected to see a comparable increase in the performance of displaced features, *i.e.*, a jump in error reduction relative to the baseline. Instead, we found just the reverse. Whereas displacement accounts for a 1.5 point gain in F (Δ_E=9.6%) with the MUC baseline features, with the beter CoNLL features, the gain due to displacement falls to 0.82 points of F (Δ_E=6.1%). While the final result with displacement (F=87.39) slightly edges out the previous high water mark of F=87.35 (Krishnan and Manning, 2005), the pattern is puzzling and not in keeping with our seven other data sets.

One possible explanations lies again with the CoNLL standard. The standard calls explicitly for inconsistent annotation of the same entity when used in different contexts. Along with place names being called MISC in hyphenated contexts (noted above), some places must be called ORG when used to refer to sports teams – except in results tables, where they are sometimes LOC. Such inconsistencies subvert the notion of long-distance dependencies by making these dependencies contradictory, thereby reducing the potential value of displacement as a means for improving performance.

7 Conclusions

Earlier in this paper, we introduced the notion of long-distance dependencies through their original codification in the context of phrase-structure grammars. By an interesting historical twist, the original solution to these grammatical long-distance effects, known as gap threading (Pereira, 1981), involved what is essentially a feature-copying operation, namely unification of constituent features. It is gratifying to note that the method presented here has illustrious predecessors.

Regarding the particular task of interest here, entity extraction, this paper conclusively shows that a simple feature-copying method provides an effective method for capturing long-distance dependencies between names. For the MUC-6 task, in particular, this error reduction is enough to lift a middle-of-the-pack performance from our baseline CRF to a level that would have placed it among the handful of top performers at the MUC-6 evaluation.

As noted, the method is also substantially more manageable than earlier approaches. It avoids the intractability of graphical models and also avoids the approximations required by methods that rely on these models. It also adds only minimal processing time at training and run times. This provides a practical alternative to the method of Krishnan and Manning, who require twelve separate training runs to create their models, and further require a time-consuming run-time process to mediate between their first and second stage CRFs.

We intend to take this work in two directions. First, we would like to get to the bottom of why the method did not do better with the CoNLL and MNET tasks. As noted earlier, our hypothesis is that we would expect greater exploitation of long-distance dependencies if we first improved the performance of the baseline CRF, especially by improving the acuity of task-related features. While it is not a key interest of ours to achieve best-in-class performance on historical evaluations, it is the case that we seek a better understanding of the range of application of the feature copying method.

Another direction of interest is to consider other problems that exhibit long-distance dependencies that might be addressed by feature copying. Word sense disambiguation is one such case, especially given Yarowsky's maxim regarding one sense per discourse, a consistency notion that seems tailor-made for treatment as long-distance dependencies (Yarowsky, 1995). Likewise, we are curious about the applicability of the method to reference resolution, another key task with long-distance effects.

Meanwhile, we believe that this method provides a practical approach for capturing long-distance effects in one of the most practical and useful application of human language technologies, entity extraction.

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

References

John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1995. Description of the Alembic system as used for MUC-6. *Pcdgs of the 6th Message Understanding Conference* (MUC-6).

Eric Brill. 1994. Some advances in rule-based part-of-speech tagging. *Pcdgs. AAAI-94*.

Razvan Bunescu and Raymond J. Mooney. 2004. Collective information extraction with relational Markov networks. *Pcdgs. of the 42nd ACL*. Barcelona.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Pcdgs. of the 43rd ACL*. Ann Arbor, MI.

Han-Shen Huang, Yu-Ming Chang, and Chun-Nan Hsu. 2007. Training conditional random fields by periodic step size adaptation for large-scale text mining. *Pcdgs. 7th Intl. Conf. on Data Mining (ICDM-2007).*

Vijay Krishnan and Christopher D. Manning. 2006. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. *Pcdgs. of the 21st COLING and 44th ACL*. Sidney.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. *Pcdgs. of the 9th EACL*. Bergen.

Fernando Pereira. 1981. Extraposition grammars. *American Jnl. of Computational Linguistics*, 4(7).

Karl-Michael Schneider. 2004. A new feature selection score for multinomial naive Bayes text classification based on KL-divergence. In *Companion to the Pcdgs. of the 42nd ACL*. Barcelona.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. *Pcdgs. of NAACL-HLT 2003*. Edmonton, CA.

Beth Sundheim, *ed.* 1995. *Pcdgs. of the 6th Message Understanding Conference (MUC-6)*. Columbia, MD.

Charles Sutton and Andrew McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. *Pcdgs. ICML Workshop on Statistical Relational Learning*.

Marc Vilain and David Day. 1996. Finite-state phrse parsing by rule sequences. *Pcdgs. of the 16th Conference on Computational Lingusitics* (COLING-96).

Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leon Peskin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Rapidly retargetable approaches to de-identification. *Journal of the Americal Medical Informatics Association*; 14(5).

Ben Wellner and Marc Vilain. (2006). Leveraging machine-readable dictionaries in discriminative sequence models. In *Pcdgs. of the 5th Language Resources and Evaluation Conf.* (LREC 2006). Genoa.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Pcdgs. Of 33rd ACL*. Cambridge, MA.