Combining Local and Non-local Information with Dual Decomposition for Named Entity Recognition from Text

Hai Leong Chieu
DSO National Laboratories
20 Science Park Drive, Singapore
Email: chaileon@dso.org.sg

Loo-Nin Teow
DSO National Laboratories
20 Science Park Drive, Singapore
Email: tloonin@dso.org.sg

***Abstract*—Named entity recognition (NER) is the task of segmenting and classifying occurrences of names in text. In NER, local contextual cues provide important evidence, but non-local information from the whole document could also prove useful: for example, it is useful to know that "Mary Kay Inc." has been mentioned in a document to classify subsequent mentions of "Mary Kay" as an organization and not as a person. Previous works for NER typically model the problem as a sequence labeling problem, coupling the predictions of neighboring words with a Markov model such as conditional random fields. We propose applying the dual decomposition approach to combine a local sentential model and a non-local label consistency model for NER. The dual decomposition approach is a fusion approach which combines two models by constraining them to agree on their predictions on the test data. Empirically, we show that this approach outperforms the local sentential models on four out of five data sets.**

I. Introduction

Named entity recognition (NER) is the task of recognizing named entities such as person, organization, or location names in texts. NER is an important task in natural language processing, as an application in itself and as a component for other tasks such as information extraction [1] or machine translation [2]. In the Sixth Message Understanding Conference (MUC6) [1], the objective of the NER task was defined as the segmentation (i.e. defining the boundaries of the named entity phrase) and classification of entities belonging to seven name-classes: PERSON, ORGANIZATION, LOCATION, DATE, TIME, MONEY and PERCENTAGE. More recently, in the CoNLL 2003 shared task [3], the NER task was defined for four name-classes: PERSON, ORGANIZATION, LOCATION, and MISCELLANEOUS, where MISCELLANEOUS names were defined as names that did not belong to the other three name-classes. We show examples of sentences annotated with named entities from five data sets in Figure 1. We will give more details on these data sets in Section IV.

NER can be formulated as a sequence labeling problem, where a sentence is treated as a sequence of words, and the objective is learn a model to predict the name-class label of each word in a sentence. Early work in the application of machine learning to NER applied hidden Markov models to the MUC6 task [4]. Since then, work on NER has been dominated by the use of discriminative models such as logistic regression [5] and conditional random fields [6].

For NER, the important cues used to classify each word to its name-class are mostly derived from surrounding words within the sentence. Examples include the case information and the surface string of the word, as well as the surrounding words, typically within a window of three to five previous and next words. However, some previous works have also shown that non-local evidence gathered from the whole document could help improve performance [7], [8]. Such evidence are usually gathered from the other occurrences of the same word in the same document. For example, it is useful to couple the prediction of mentions of "Mary Kay Inc." with subsequent mentions of "Mary Kay" so that both mentions are labeled as ORGANIZATION and not as PERSON. It is hence interesting to define NER as a structured prediction problem where the prediction of the labels of each word is not only coupled with the prediction of its surrounding words, but also with the labels of other occurrences of the same word. Finkel et al. [7] illustrated this *label consistency* problem with the example in Figure 2. However, a graphical model that couples both local and non-local label prediction would result in an intractable inference problem [9]. Finkel et al. [7] solved this by defining a factored sequence model, which combines a local sentential model and a label consistency model with the product of the two models. For the label consistency model, they derived penalties scores for inconsistencies from the training data. They performed approximate inference in the factored model using Gibbs sampling.

Recently, the *dual decomposition* approach has been used for natural language processing applications such as parsing [10], [11]. Dual decomposition first breaks down an intractable inference problem into tractable sub-problems, and then constrains the models of the sub-problems to agree on the test data. This is formulated as an optimization problem that optimizes the sum of the objectives of the sub-models, subject to constraints that the sub-models agree on the data. The optimization problem can be solved in the dual, using a linear programming relaxation. Despite the relaxation, it has been shown that dual decomposition often recover the provably

Example from the CoNLL 2003 data set (ENG):
The European Commission [ORG] said on Thursday it disagreed with German [MISC] advice to consumers to shun British [MISC] lamb until scientists determine whether mad cow disease can be transmitted to sheep. Germany [LOC]'s representative to the European Union [ORG]'s veterinary committee Werner Zwingmann [PER] said on Wednesday consumers should buy sheepmeat from countries other than Britain [LOC] until the scientific advice was clearer.

Example from the CoNLL 2003 data set (DEU):
Diskussionen über Asylbewerber-, Aussiedler- und Einwanderungsfragen haben in Deutschland [LOC] in den letzten Monaten Politik und Medien beherrscht. Es ist deswegen momentan nicht schwer, mit der Publikation eines Buches über Flüchtlinge und Vertriebene in Westdeutschland [LOC] von 1945 bis 1990 Aufmerksamkeit zu erregen.

Example from the MUC6 data set (MUC6):
Mary Kay [ORG] Names Vice Chairman ... Richard C. Bartlett [PER] was named to the newly created position of vice chairman of Mary Kay Corp. [ORG], a privately held cosmetics company. Mr. Bartlett [PER] was previously president and chief operating officer of Mary Kay Cosmetics Inc. [ORG], the company's operating subsidiary.

Example from the twitter data set (TWIT):
Pxleyes [ORG] Top 50 Photography Contest Pictures of August 2010 ... http://bit.ly/bgCyZ0#photography

Example from the maritime data set (RECAAP):
Among the incidents reported in April 2010 were two incidents of hijacking of tug boats. They were the hijacking of tug boat, PU 2007 [VES] in the South China Sea on 19 April 2010, and the hijacking of tug boat Atlantic 3 [VES] at about 11 nm east of Pulau Bintan, Indonesia on 27 April 2010.

Fig. 1. Examples of sentences labeled with named entities belonging to the classes PERSON (PER), LOCATION (LOC), ORGANIZATION (ORG), MISCELLANEOUS (MISC) and VESSELS (VES). The named entities are underlined and annotated with their name-classes in square brackets in the above examples. The name-classes defined for each data set may be different. For example. in the RECAAP data set, only vessel names are annotated.

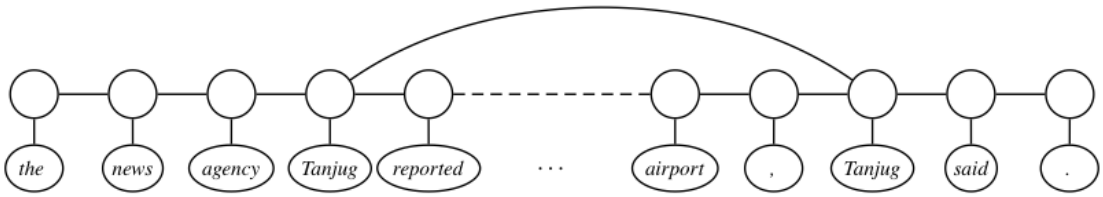

Fig. 2. This figure is used in [7] to illustrate the label consistency problem which couples the predictions of labels that are not adjacent to one another. Without modeling the edge between the two occurrences of "Tanjung", the predictions of the two occurrences are decoupled given the labels of the other words between the two occurrences.

optimal solution of the optimization problem [10]–[12].

In this paper, we propose learning two conditional random fields (CRF) models from the training data: a typical sentence conditional random field (SCRF) model that models the sentence as a sequence of labels, and a second model, a word conditional random field (WCRF) model, that ignores sentence structure, and chains the different occurrences of the same word as a sequence of labels. After training the CRF on the training data (manually annotated with named entities), the models learned can then be applied to the test data to predict the name-classes of each word in the test data. We apply the dual decomposition approach during this test phase to enforce agreement between these two models on the test data. The rest of the paper is organized as follows. In Section II, we give a brief review of conditional random fields (CRF) and describe the SCRF and WCRF models we defined for the NER task. In Section III, we describe the dual decomposition approach and its application to combine SCRF and WCRF. In Section IV, we show that the dual decomposition approach outperforms the SCRF model in four out of five data sets, and discuss why the approach fails to improve performance on the fifth data set. We describe related work in Section V, and we conclude in Section VI.

II. Conditional Random Fields

In this section, we first give a brief review of conditional random fields (CRF) [13]. Conditional random fields [13] are discriminative, undirected graphical models. They have been shown to perform well in a variety of tasks including part-of-speech tagging [13], named entity recognition [6], shallow-parsing [14] and object recognition in computer vision [15]. In this paper, we apply conditional random fields to the NER

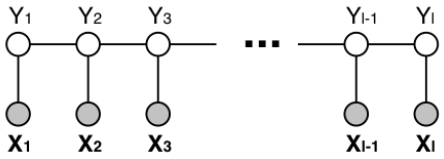

Fig. 3. Graphical representation of a linear chain conditional random fields. Shaded nodes are observed, and the graph G in Definition 1 is the subgraph induced by the nodes $\mathbf{Y}$.

problem as a sequence labeling problem.

The sequence labeling problem is the assignment of labels to a sequence of observations. For NER, each sentence is a sequence of words, and the objective of the NER problem is to assign named entity labels to each word in the sequence. Given a sequence, we denote $\mathbf{X}$ as the vector of features that can be extracted from the observed sentence, and $\mathbf{Y}$ as the corresponding sequence of labels for the sequence (see Figure 3). Each component Y_i of $\mathbf{Y}$ range over an alphabet *Y*. In general, a CRF is defined as follows [13]:

Definition: Let $G = (\mathbf{V}, \mathbf{E})$ be a graph such that $\mathbf{Y} = (Y_v)_{v \in \mathbf{V}}$, so that $\mathbf{Y}$ is indexed by the vertices of G. Then $(\mathbf{X}, \mathbf{Y})$ is a conditional random field in case, when conditioned on $\mathbf{X}$, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|\mathbf{X}, \mathbf{Y}, w \sim v) = p(Y_v|\mathbf{X}, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

By the fundamental theorem of random fields [16], the general form of the joint distribution of the labeled sequence $\mathbf{Y}$ given $\mathbf{X}$ has the form:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\sum_{c \in C} \Phi(\mathbf{y_c}, \mathbf{x})), \quad (1)$$

where C is the set of cliques in the graph G, and $Z(\mathbf{x})$ is a normalization factor.

In applications of the CRF, the $\mathbf{X}$ are usually observed in the data, and the task involves the prediction of the labels $\mathbf{Y}$, given the observations $\mathbf{X}$. For example, for NER, the $\mathbf{X}$ are features derived from the surface strings of words in a sentence, and the sequence of labels $\mathbf{Y}$ is the sequence of named entity labels assigned to each word in the sentence.

In the case of a linear chain (shown in Figure 3), the joint distribution can be expressed as [13]

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left[\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_{\mathbf{e}}, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_{\mathbf{v}}, \mathbf{x}) \right], \quad (2)$$

where $\mathbf{x}$ is a data sequence, $\mathbf{y}$ a label sequence and $y|_s$ is the set of components of $\mathbf{y}$ associated with the vertices in the subgraph S. The functions f_k and g_k are features: f_k are features on the edges of the form (Y_i, Y_{i+1}) and g_k features on the edges of the form $(\mathbf{X_i}, Y_i)$ in the linear chain. To simplify notation, when it is not necessary to distinguish between f_k and g_k, we write

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})), \quad (3)$$

where $F(\mathbf{y}, \mathbf{x})$ is the global feature vector for the input sequence $\mathbf{x}$ and label sequence $\mathbf{y}$, comprising of the f_k's and the g_k's.

The parameter estimation problem is to determine, from the training data $D = \{(\mathbf{x^{(j)}}, \mathbf{y^{(j)}})\}_{j=1..N}$, the parameters in Λ. We determine Λ by maximizing the log-likelihood of the training data:

$$L_\Lambda = \sum_j [\mathbf{\Lambda}.\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - \log Z_\Lambda(\mathbf{x^{(j)}})]. \quad (4)$$

It is often useful to define a Gaussian prior over the parameters to avoid overfitting (a process that is sometimes called regularization), which changes the above objective function into

$$L_\Lambda = \sum_j [\mathbf{\Lambda}.\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - \log Z_\Lambda(\mathbf{x^{(j)}})] - \frac{\|\mathbf{\Lambda}\|^2}{2\sigma^2}. \quad (5)$$

We use a gradient based algorithm for maximizing the log likelihood, which requires the calculation of the gradient of the regularized log-likelihood [14]:

$$\nabla L_\Lambda = \sum_j [\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - E_{p_\Lambda(\mathbf{Y}|\mathbf{x^{(j)}})} F(\mathbf{Y}, \mathbf{x^{(j)}})] - \frac{\Lambda}{\sigma^2}. \quad (6)$$

The above gradient term requires the calculation for each sequence $\mathbf{X}$ of the expected feature values over all possible Y over the entire sequence $\mathbf{X}$. For the linear chain, this can be done efficiently by the forward backward algorithm [17]. The gradient based approach we used the L-BFGS for Quasi Newton unconstrained minimization, implemented in the Stanford NLP toolkit [18]. In all our experiments, we set $\frac{1}{2\sigma^2}$ to 0.1.

A. Conditional Random Fields for NER

We applied two conditional random fields models to address the NER problem. The first model, called the Sentence Conditional Random Field (SCRF) model, is the conventional application of conditional random fields to NER as a sequence labeling problem, where each sentence is treated as a sequence. The second model, called the Word Conditional Random Field (WCRF) ignores the sentence structure and chains identical words in a linear chain.

In SCRF, each sequence $\mathbf{x^{(j)}}$ represents a sentence, and each $\mathbf{y^{(j)}}$ the sequence of labels assigned to the words in the sentence. For NER, the alphabet *Y* of the sequence of labels is the set of possible name-classes. For example, for the CoNLL 2003 shared task [3], the alphabet *Y* consists of the name-classes PER, ORG, LOC, MISC, and a not-a-name class. The features f_k and g_k are defined on single labels and pairs of adjacent labels respectively. For NER,

TABLE I
FEATURES USED FOR NER FOR EACH WORD OF A SENTENCE. THE LAST THREE FEATURE TYPES ARE OF TYPE g_k, WHILE THE OTHER FEATURES TYPES ARE OF TYPE f_k, DEFINED IN EQUATION 2.

Feature type	Features
Word features	all words in a window from the previous 4 to the next 4 words
Word bigrams	bigrams composed of previous, current and next words
Word shape features	word shapes of previous, current and next words
Word shape bigrams	bigrams of word shapes of previous, current and next words
Character n-grams	character n-grams of the current word
Other occurrences	indicative feature if other occurrences of the same word starts with an uppercase character
Gazetteer features	indicative feature of the presence of phrases found in gazetteers
Pairwise label feature (g_k)	feature of current label with next label
Pairwise word features (g_k)	bigram feature of word shapes of the current and next word
Pairwise bigram of word shape (g_k)	bigram feature of word shapes of the current and next word

we used standard features that have been found useful for NER in previous works, such as word strings and word shape features. Word shape features [19] are features that maps each word onto equivalence classes that encodes attributes such as length, capitalization, and numerals. For instance, ex-England would be represented as "xx- Xxxx", and "37-run" would be represented as "dd-xxx". As gazetteers (list of names) have been shown to be useful for NER, we built gazetteers for each name-class of interest by downloading from resources on the Web. We give more details about how the gazetteers are constructed in our experiments in Section IV. The full list of features used are shown in Table I.

For WCRF, the linear chains are constructed by chaining together words that are written in the same way (identical spelling and word shape) in the entire test set. The features used for the WCRF are the same as those used for the SCRF. We illustrate the construction of the SCRF and WCRF models for two sentences in Figure 4. For words occurring only once in the test set, we could model them in WCRF as singletons (graphs with only one node). However, we found that this usually results in worse performance for the dual decomposition approach. This is natural as words modeled as singletons in the WCRF would not add any useful complementary information to the SCRF. Hence, in all our experiments, we discard singletons in the WCRF, both during training and testing. After discarding singletons, the WCRF does not predict the labels of the entire test set, but only the labels of words that occurred at least twice. We could have use the WCRF alone to label the entire test set if we do not discard singletons, but in our experiments, the performance of WCRF is always worse than the SCRF.

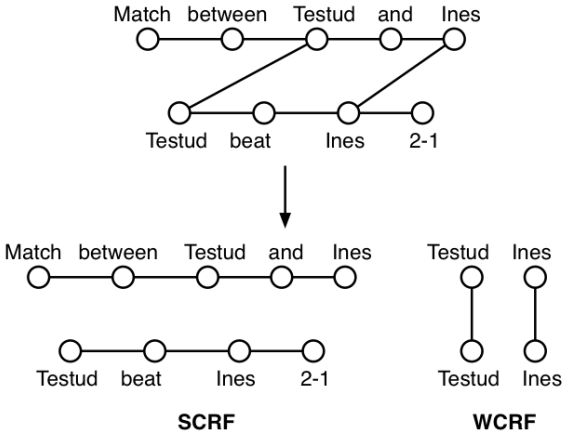

Fig. 4. Illustration of the decomposition process for NER. Inference is intractable for the structure at the top of the figure, and hence we decompose it into the two sub structures below which are tractable. The structure on the left is the sentence based structure for the SCRF, while the structure on the right is the word based structure for the WCRF.

III. DUAL DECOMPOSITION

We used the dual decomposition approach to combine the outputs of SCRF and WCRF. The dual decomposition approach [10] addresses structured prediction problems for which inference in the graphical structures may be intractable. In such cases, Rush et al. [10] proposed to decompose the intractable structure into tractable sub-structures, and to force the models on the sub-structures to agree on their predictions on the test data. We apply the dual decomposition approach to constraint the agreement of SCRF and WCRF described in the previous section. Our models WCRF and SCRF will have separate objective functions $f(\mathbf{Y})$ and $g(\mathbf{Z})$ respectively, where $\mathbf{Y}$ consists of chains of labels where each chain is a sentence, while $\mathbf{Z}$ consists of chains of labels where each chain connect occurrences of the same word. (To simplify notations, we ignore the dependency of the functions on the entire set of test observations, $\mathbf{x}$). The joint objective function is hence defined as

$$\begin{aligned} &\arg\max_{\mathbf{y},\mathbf{z}} f(\mathbf{y}) + g(\mathbf{z}), \\ &\text{such that for each word } w, y_w = z_w. \end{aligned} \tag{7}$$

This constrained optimization problem can then be converted to an unconstrained optimization problem by introducing Lagrangian multipliers for the constraints [12],

$$L(\mathbf{u}, \mathbf{y}, \mathbf{z}) = f(\mathbf{y}) + g(\mathbf{z}) + \sum_w u(w)(y_w - z_w). \tag{8}$$

The dual problem is optimizing for $\min_{\mathbf{u}} L(\mathbf{u})$, where

$$L(\mathbf{u}) = \max_{\mathbf{z},\mathbf{y}} L(\mathbf{u}, \mathbf{y}, \mathbf{z}). \tag{9}$$

This objective can be optimized using gradient descent, resulting in Algorithm 1. In this algorithm, we define our strategy

Algorithm 1 Algorithm for gradient descent for optimization of the dual decomposition objective.

Input: The predictions of the SCRF, $\mathbf{y}^0$, and the WCRF, $\mathbf{z}^0$ on the test data.
Output: Final predictions of the labels of the test data.
Algorithm:

1) Set $k \leftarrow 0$.
2) Initialize $\mathbf{u}^0$ to a vector of zeros, $\alpha \leftarrow 0.5, \delta \leftarrow 0$
3) **While** not converged
   a) $\mathbf{y}^k \leftarrow \arg\max_{\mathbf{y}} \sum_w (f(y_w) - u^k_{w,y_w})$
   b) $\mathbf{z}^k \leftarrow \arg\max_{\mathbf{z}} \sum_w (g(z_w) + u^k_{w,z_w})$
   c) if $\mathbf{y}^k = \mathbf{z}^k$, then converged = true
   d) for each word w,

$$\begin{aligned} u^{k+1}_{y,w} &\leftarrow u^{k+1}_{y,w} + \alpha_k \text{ and} \\ u^{k+1}_{z,w} &\leftarrow u^{k+1}_{z,w} - \alpha_k \end{aligned}$$

   e) $k \leftarrow k + 1$
   f) if $L(\mathbf{u}^{k+1}) > L(\mathbf{u}^k), \delta \leftarrow \delta + 1.0$
   g) $\alpha \leftarrow \alpha/(1 + \delta)$
   h) if $\alpha < \epsilon$, converged = true

   end while
4) Output $\mathbf{y}^k$.

of gradually decreasing the value of α, in a way similar to those proposed in [10], [11].

If Algorithm 1 terminates with $\mathbf{y}^k = \mathbf{z}^k$, it can be shown that this solution *exactly* solves the primal problem in Equation 7 [11]. In our experiments, we constrain the SCRF and WCRF to agree over the entire test set. In all our experiments, we do not achieve the $\mathbf{y}^k = \mathbf{z}^k$ condition, and we terminate the algorithm when $\alpha < \epsilon$, and where they disagree, we output the results of the SCRF (i.e., the final $\mathbf{y}^k$ in Algorithm 1). In our experiments, we set ϵ to be 10^{-6}.

IV. Experimental Results

In this section, we describe and discuss experimental results comparing the dual decomposition approach with the sentence based CRF (SCRF). We first describe the five data sets used in our experiments:

- **The CoNLL-2003 English data (ENG)** consists of a training set and two test sets, a developmental test set *testa*, and an official test set, *testb* [3]. There are four name-classes in this data set: PERSON (PER), LOCATION (LOC), ORGANIZATION (ORG), and MISCELLANEOUS (MISC). To compare against previous published results, we evaluate on the official test set, *testb*.
- **The CoNLL 2003 German data (DEU)** consists of a training set and two test sets, *testa* and *testb*. It has the same name-classes as ENG [3]. Similarly, we evaluate on the DEU *testb* data set to be able to compare against previous results.
- **The MUC6 data (MUC6)** is the benchmark data set for the *Sixth Message Understanding Conferences* [1]. We use three of the MUC6 name-classes: PER, LOC and ORG. We use the MUC6 training set for training, and combine the dry-run and formal test sets for testing.
- **The Twitter data (TWIT)** is the Twitter NER data set used in [20]. The data set has ten name-classes, but as a few name-classes are rather rare, we only evaluate on three of the name-classes: PER, LOC and ORG. We use the last 1,000 tweets as test data and the rest as training data.
- **The RECAAP data set (RECAAP)** is a collection of incident and quarterly reports downloaded from the website for *The Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia* [21]. The data set consists of 97 documents with a total of 324,108 tokens, labeled with 2,449 vessel names (VES). We use the first 60 documents for training and the remaining 37 documents for testing.

TABLE II
Number of documents or tweets in the training and test data for each data set.

Data	Training Data	Test Data
ENG	946	231
DEU	553	155
MUC6	318	60
TWIT	1394	1000
RECAAP	60	37

Each data set consists of a training data and a test data. The training data is used to train the SCRF and the WCRF models. The test data is used for evaluation purposes only. The dual decomposition algorithm is applied on the test data to combine the outputs of the SCRF and the WCRF models.

For the gazetteer feature, we build the gazetteers from resources on the Web for each name-class. The gazetteer for the not-a-name class for each data set consists of the high frequency words in the data set, verbs and capitalized words (such as weekdays for English or common nouns for German). We give the detailed sources of the gazetteers in Table III.

A. Evaluation Measure

We evaluate all our results using the official scorer provided by the CoNLL 2003 shared task [3]. We refer to the manually annotated data as the *ground truth*, and we measure the system's performance by comparing its predictions on the test data against the ground truth. The performance is measured using the F_1-measure. More generally, the F_β-measure [22] is defined as

$$F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \tag{10}$$

where precision is the number of correctly predicted positives (true positives) over the total number of predicted positives, and recall is the number of correctly predicted positives (true positives) over the total number of positives in the ground truth. For NER, a named entity prediction is considered correct

TABLE III

GAZETTEERS USED FOR THE FIVE DATA SETS. THE GAZETTEERS ARE CONSTRUCTED FROM ONLINE RESOURCES, AS WELL AS OFFICIAL GAZETTEERS PROVIDED DURING THE CONLL 2003 SHARED TASK. IN OUR WORK, WE HAVE ORGANIZED THE GAZETTEERS SUCH THAT WE HAVE ONE GAZETTEER PER NAME-CLASS. IN GENERAL, GAZETTEERS COULD SIMPLY BE LISTS OF PHRASES THAT COULD BE CORRELATED WITH THE NAME-CLASSES OF INTEREST.

Data sets	Gazetteers
ENG, MUC6	**PER:** list of top 1000 people's first and last names from the US census bureau. **ORG:** Forbes' 500 list in 2003, list of football clubs. **LOC:** list of countries, capitals. **O:** 500 most frequent words in the ENG corpus, Monday to Sunday, January to December, prepositions and verbs.
ENG	**MISC:** list of nationalities.
DEU	**PER, ORG, LOC, MISC:** name lists provided at CoNLL 2003 shared task. **O:** 100 most frequent words in the DEU corpus, and 50 common German nouns.
TWIT	**PER:** The PER gazetteer for ENG, plus a list of people names from Wikipedia. **ORG:** The ORG gazetteer for ENG, plus the relevant lists from twitter_nlp website. **LOC:** The LOC gazetteer for ENG, plus the relevant lists from twitter_nlp website. **O:** The O gazetteer for ENG, plus words occurring more than 5 times in the TWIT data set and lists from the twitter_nlp website.
RECAAP	**VES:** top 500 ship names from a maritime database. **O:** the O gazetteer for ENG, plus 100 most frequent words in the RECAAP data.

only if it is an exact match of the corresponding entity in the ground truth (segmented at the same boundaries), and predicted to be of the same name-class.

Our system, as it is implemented, would be unable to segment consecutive entities of the same name-class. For example, in the phrase

> Benjamin Netanyahu vowed to retain the Golan Heights [LOC] Israel [LOC] captured from Syria [LOC],

our system can only segment

> Golan Heights Israel [LOC]

as a single named entity. For the CoNLL task, such a segmentation will be scored as wrong against the ground truth that is annotated in the ENG data. One common way to address this problem is to augment the alphabet of the labels with classes to mark the beginning of a name-class [23]. For example, the PER class will be broken into two separate classes: B-PER for the first word of a PER entity, and I-PER for subsequent words. However, this will double the number of name-classes, making inference in the linear chain more expensive. (Inference is quadratic in the number of classes.) As such cases of continuous entities of the same class are very rare, we used the simple label representation without distinguishing between first and subsequent words.

B. Results and Discussion

We tabulate the results comparing SCRF with the dual decomposition approach on the five data sets in the Tables IV to VIII. First, we compare our results with previous published results on the ENG and DEU *testb* data set, shown in Table IV for ENG and Table V for DEU. The dual decomposition approach achieves 87.89% on the ENG *testb* data, and 71.34% on the DEU *testb* data. This is close to the results of the top system in the CoNLL 2003 shared task, which achieved 88.76% and 72.41% on the ENG and the DEU *testb* data respectively [24]. As the CoNLL 2003 shared task is a competition and competitors are allowed to use external resources, the top system in CoNLL 2003 shared task [24] used a classifier combination approach that combines NER classifiers that are trained with additional more generally annotated data. Our results also compare favorably with the results of Finkel et al. [7], who reported an improvement from 85.51% to 86.86% on the ENG *testb* data set with their factored model.

From the tables, we see that the dual decomposition approach outperforms SCRF on all data sets except TWIT. We postulate that in the tweets in the TWIT data, the local consistency information may be less important. Tweets are limited in length, and within the same tweet, the same word is unlikely to occur twice. Although we chain all identical words in the entire test set, the phenomenon of label consistency for identical words may not be as strongly manifest in tweets as it is in newswire articles.

V. RELATED WORK

The NER task is an important task in natural language processing, and the availability of benchmark NER data sets such as the MUC6 data [1] and the CoNLL 2003 shared task data [3] have resulted in a large number of recent papers published on this task. In MUC6, BBN's hidden Markov model (HMM) based IdentiFinder [4] achieved remarkably good performance. The Seventh Message Understanding Conference (MUC7) [25] has also seen hybrids of statistical NERs and hand-coded systems [5], [26], notably Mikheev et al. [26], which achieved the best performance of 93.39% on the official NER test data. MENE (Maximum Entropy Named Entity) [5] was combined with Proteus (a hand-coded system), and came in fourth among all MUC- 7 participants. MENE without Proteus, however, did not do very well and only achieved an F_1-measure of 84.22% [5].

Since MUC6 and MUC7, the NER task are often defined to exclude the extraction of numerical entities such as DATE and PERCENTAGE. The CoNLL 2003 shared task [3], for example, defined the NER task as the extraction of four name-classes, namely PERSON, ORGANIZATION, LOCATION

TABLE IV
PRECISION (PRE), RECALL (REC) AND F_1-MEASURE (F_1) FOR EACH NAME-CLASS FOR THE ENG *testb* DATA SET.

	SCRF			Dual Decomposition		
Data	Pre	Rec	F_1	Pre	Rec	F_1
LOC	90.89	90.67	90.78	90.54	90.97	90.76
MISC	81.71	77.34	79.47	81.97	79.37	80.65
ORG	82.86	80.25	81.53	83.65	82.49	83.06
PER	90.50	91.34	90.92	91.50	94.56	93.00
All	87.37	86.16	86.76	87.81	88.08	**87.94**

TABLE V
PRECISION (PRE), RECALL (REC) AND F_1-MEASURE (F_1) FOR EACH NAME-CLASS FOR THE DEU *testb* DATA SET.

	SCRF			Dual Decomposition		
Data	Pre	Rec	F_1	Pre	Rec	F_1
LOC	80.28	67.25	73.19	80.94	68.50	74.20
MISC	75.98	47.26	58.27	76.59	47.87	58.91
ORG	78.55	55.89	65.31	78.84	56.40	65.76
PER	91.55	68.15	78.14	88.51	70.41	78.43
All	82.97	61.55	70.68	82.46	62.87	**71.34**

TABLE VI
PRECISION (PRE), RECALL (REC) AND F_1-MEASURE (F_1) FOR EACH NAME-CLASS FOR THE MUC6 DATA SET.

	SCRF			Dual Decomposition		
Data	Pre	Rec	F_1	Pre	Rec	F_1
LOC	80.08	80.69	80.38	80.00	80.31	80.15
ORG	88.21	71.37	78.90	87.99	74.25	80.54
PER	91.50	94.65	93.05	92.49	96.56	94.48
All	88.00	79.53	83.55	88.22	81.66	**84.81**

TABLE VII
PRECISION (PRE), RECALL (REC) AND F_1-MEASURE (F_1) FOR EACH NAME-CLASS FOR THE TWIT DATA SET.

	SCRF			Dual Decomposition		
Data	Pre	Rec	F_1	Pre	Rec	F_1
ORG	92.00	34.33	50.00	96.30	38.81	55.32
LOC	82.69	34.68	48.86	85.71	33.87	48.55
PER	71.57	43.71	54.28	70.00	41.92	52.43
All	77.65	38.83	**51.77**	78.41	38.55	51.69

TABLE VIII
PRECISION (PRE), RECALL (REC) AND F_1-MEASURE (F_1) FOR EACH NAME-CLASS FOR THE RECAAP DATA SET.

	SCRF			Dual Decomposition		
Data	Pre	Rec	F_1	Pre	Rec	F_1
VES	88.78	80.00	84.16	88.31	81.89	**84.98**
All	88.78	80.00	84.16	88.31	81.89	**84.98**

and MISCELLANEOUS. The addition of the MISCELLANEOUS class reflects the fact that there are names which do not belong to the conventional name-classes, and which could be important in practice. Sekine et al. [27] went on to define an extended named entity hierarchy comprising of 150 named entity classes. NER has also found applications in the medical domain, where NER is defined to be the task of recognizing names belonging to classes such as protein, DNA, RNA, cell line and cell type [28].

Most published machine learning NER systems only use local contextual information as features, with a few notable exceptions. Mikheev et al. [26] made use of information from the whole document, using a hybrid of hand-coded rules and machine learning methods. Another attempt at using label consistency information can be found in [5], where Borthwick used an additional maximum entropy classifier that tries to correct mistakes by using reference resolution (finding words that co-refer to the same entity). Chieu and Ng [8] defined additional features for each word taken that are derived from other occurrences of the same word in the document. This approach has the advantage of allowing the training procedure to automatically learn good weightings for these global features relative to the local ones. However, they did not couple the prediction of identical words. Finkel et al. [7] coupled the prediction of identical words by combining two sequence model into a factored sequence model. However, as they noted in their paper, their factored model generates the label sequence twice. In this paper, we formulated a dual decomposition approach that combines a local and a non-local model, and we showed in our experiments that we achieved better results than [7] on the CoNLL English *testb* data set.

The dual decomposition [29] approach has recently been applied to a number of natural language processing tasks such as dependency parsing [12]. In many common natural language problems, richer models tend to perform better but the decoding process for rich models tend to be expensive or intractable. Dual decomposition can be applied in such cases as a fast approximate algorithm that could provide certificates of optimality. The dual decomposition approach consists of first decomposing a complicated optimization problem into tractable sub problems, and then combining the solutions of the sub problems iteratively into a "better" solution. In natural language processing, it has recently been applied to combining parsing and part-of-speech tagging [10], syntactic machine translation [30], symmetric HMM alignment [31], and high-order non-projective dependency parsing [11]. In this paper, we apply the dual decomposition approach for decoding in the NER problem, and we show that it outperforms the baseline sentence-based conditional random field model.

VI. CONCLUSION

The use of graphical models in structured prediction problems have shown to be effective in many areas. One problem with learning using structured prediction models is that inference in the graphical model becomes easily intractable when the graph structure contains loops [9]. The dual decomposition

approach has been shown to be effective as an approximate approach for structured prediction in many problems [10]–[12]. In this paper, we applied the dual decomposition approach to combine local and non-local information for the NER task. Empirically, we show that this approach outperforms the sentence based NER system in four out of five data sets. In the TWIT data set where dual decomposition approach fails to improve performance, we postulate that the reason is due to the fact that label consistency information is less important for tweets than for other kinds of texts.

REFERENCES

[1] United States Defense Advanced Research Projects Agency, Information Technology Office, *Proceedings of the Sixth Message Understanding Conference*. San Francisco, CA: Morgan Kaufmann, 1995.

[2] U. Hermjakob, K. Knight, and H. Daumé III, "Name translation in statistical machine translation - learning when to transliterate," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 389–397. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1045

[3] K. S. E. F. Tjong and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of CoNLL-2003*, W. Daelemans and M. Osborne, Eds. Edmonton, Canada, 2003, pp. 142–147.

[4] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder," in *Proceedings of the fifth conference on Applied natural language processing*, ser. ANLC '97. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997, pp. 194–201.

[5] A. E. Borthwick, "A maximum entropy approach to named entity recognition," Ph.D. dissertation, New York, NY, USA, 1999, aAI9945252.

[6] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields," in *Proceedings of Conference on Computational Natural Language Learning*, 2003.

[7] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2005, pp. 363–370.

[8] H. L. Chieu and H. T. Ng, "Named entity recognition: A maximum entropy approach using global information," in *COLING'02: Proceedings of the 19th international conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 1–7.

[9] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[10] A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola, "On dual decomposition and linear programming relaxations for natural language processing," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, October 2010, pp. 1–11.

[11] T. Koo, A. M. Rush, M. Collins, T. Jaakkola, and D. Sontag, "Dual decomposition for parsing with non-projective head automata," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, October 2010, pp. 1288–1298.

[12] M. Collins and A. M. Rush, "Dual decomposition for natural language processing," in *ACL (Tutorial Abstracts)*, 2011, p. 6.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[14] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 282–289.

[15] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *NIPS'04: Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 2005, pp. 1097–1104.

[16] J. Besag, "Patial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974.

[17] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.

[18] C. Manning and D. Klein, "Optimization, maxent models, and conditional estimation without magic," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 8–8.

[19] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition: From syntax to the web," in *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, N. Collier, P. Ruch, and A. Nazarenko, Eds. Geneva, Switzerland: COLING, August 28th and 29th 2004, pp. 91–94.

[20] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1524–1534.

[21] RECAAP, "The Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia (ReCAAP); http://www.recaap.org/," 2011.

[22] C. van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[23] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum entropy approach," in *CoNLL'03: Proceedings of the Seventh Conference on Natural Language Learning (Shared Task Paper)*, W. Daelemans and M. Osborne, Eds. Edmonton, Canada, 2003, pp. 160–163. [Online]. Available: conll03.pdf

[24] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proceedings of CoNLL-2003*, W. Daelemans and M. Osborne, Eds. Edmonton, Canada, 2003, pp. 168–171.

[25] N. Chinchor, *Proceedings of the Seventh Message Understanding Conference*, E. Voorhees, Ed. San Francisco, CA: Science Applications International Corporation (SAIC), 1998.

[26] A. Mikheev, C. Grover, and M. Moens, "Description of the ltg system used for muc-7," in *In Proceedings of 7th Message Understanding Conference (MUC-7*, 1998.

[27] S. Sekine, K. Sudo, and C. Nobata, "Extended named entity hierarchy," in *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, M. G. Rodríguez and C. P. S. Araujo, Eds., Canary Islands, Spain, May 2002, pp. 1818–1824.

[28] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at jnlpba," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, ser. JNLPBA '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, pp. 70–75.

[29] N. Komodakis, N. Paragios, and G. Tziritas, "Mrf energy minimization and beyond via dual decomposition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 3, pp. 531 –552, march 2011.

[30] A. M. Rush and M. Collins, "Exact decoding of syntactic translation models through lagrangian relaxation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 72–82. [Online]. Available: http://www.aclweb.org/anthology/P11-1008

[31] J. DeNero and K. Macherey, "Model-based aligner combination using dual decomposition," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 420–429. [Online]. Available: http://www.aclweb.org/anthology/P11-1043