Part of Speech Tagging for French Social Media Data

Farhad Nooralahzadeh
École Polytechnique
de Montréal
Montréal, PQ, Canada
nooralahzadeh
@gmail.com

Caroline Brun
Xerox Research Centre
Europe
Meylan, France
caroline.brun
@xrce.xerox.com

Claude Roux
Xerox Research Centre
Europe
Meylan, France
claude.roux
@xrce.xerox.com

Abstract

In the context of Social Media Analytics, Natural Language Processing tools face new challenges on on-line conversational text, such as microblogs, chat, or text messages, because of the specificity of the language used in these channels. This work addresses the problem of Part-Of-Speech tagging (initially for French but also for English) on noisy language usage from the popular social media services like Twitter, Facebook and forums. We employ a linear-chain conditional random fields (CRFs) model, enriched with several morphological, orthographic, lexical and large-scale word clustering features. Our experiments used different feature configurations to train the model. We achieved a higher tagging performance with these features, compared to baseline results on French social media bank. Moreover, experiments on English social media content show that our model improves over previous works on these data.

1 Introduction

There are many challenges inherent to applying standard natural language analysis techniques to social media. On-line conversational texts, such as tweets are quite challenging for text mining tools, and in particular for opinion mining, as they contain very little contextual information and assume too much implicit knowledge. They expose much more language variation and tend to be less grammatical than regular texts such as news articles or books. Furthermore, they contain unusual capitalization, and make frequent use of emoticons, abbreviations and hash-tags, which can form an important part of their inner meaning (Maynard et al., 2012). Conventional natural language processing tools for regular texts have achieved reasonably high accuracy thanks to machine learning techniques on large annotated data set. However, "off the shelf" language processing systems fail to work on social media data and their performance on this domain degrade very fast. For example, in English Part-Of-Speech tagging, the accuracy of the Stanford tagger (Toutanova et al., 2003) falls from 97% on Wall Street Journal text to 85% accuracy on Twitter (Gimpel et al., 2011), similarly the MElt POS tagger (Denis and Sagot, 2012) drops from 97.7% on the French Treebank (called the FTB-UC by (Candito and Crabbé, 2009)) to 85.2% on on-line conversational texts (Seddah et al., 2012). In Named Entity Recognition, the CoNLL-trained Stanford recognizer achieves 44% F-measure (Ritter et al., 2011), down from 86% on the CoNLL test set (Finkel et al., 2005); regarding parsing, see for example (Foster et al., 2011; Seddah et al., 2012), poor performances have been reported for different state-of-the-art parsers applied to English and French social media content.

The main objective of this work is to implement a dedicated Part-Of-Speech (POS) tagger for French social media content such as Twitter, Facebook, blogs, forums and customer reviews. We used the first user-generated content resource for French presented by Seddah et al. (2012), which contains a fine-grained tag set and has been extracted from various social media contents. We have designed and implemented a POS tagger considering one of the well-known *discriminative* type of sequence-based methods; Conditional Random Fields (CRF) (Lafferty et al., 2001). To deal with sparsity and unknown

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1764–1772, Dublin, Ireland, August 23-29 2014.

words, we have applied unsupervised techniques to enrich the feature set. Finally, we have evaluated our tagger performance with different configurations on annotated corpora from French social media.

We will first present related work in Part-Of-Speech tagging (Section 2) on noisy data like social media content. In Section 3, the annotated dataset and its characteristics (e.g., tag set) are described. Section 4 presents the result of applying the MElt POS tagger to user generated text as our baseline (Seddah et al., 2012). In Section 5, we explain how we design and implement our POS tagger. Section 6 is devoted to experiments and performance of our tagger. Section 7 describes the evaluation of the new tagger on English social media texts. Conclusion and future work are given in Section 8.

2 Related work

Online conversational texts, typified by micro-blogs, chat, and text messages, are a challenge for natural language processing. Unlike the highly edited genres for which conventional NLP tools have been developed, conversational texts contain many non-standard lexical items and syntactic patterns. These are the result of unintentional errors, dialectal variation, conversational ellipsis, topic diversity, and creative use of language and orthography (Eisenstein, 2013)

The language technology research community proposes two approaches to deal with noisy texts, namely normalization and domain adaptation, which are briefly described here.

2.1 Normalization

One way to deal with ill-formed language is to turn it into a well-formed language as a pre-processing task: "normalizing" social media or SMS messages to better conform to the language that the technology expects. For example, (Han and Baldwin, 2011) propose the lexical normalization of short text messages, such as tweets, based on string and distributional similarity. They describe a method to identify and normalize ill-formed words. Word similarity and context are exploited to select the best candidate for noisy tokens.

2.2 Domain adaptation

The other approach is instead to adapt the tools to fit the text. A series of papers has followed the mold of "NLP for Twitter," including POS tagging (Gimpel et al., 2011; Owoputi et al., 2013), named entity recognition (Finin et al., 2010; Ritter et al., 2011; Xiaohua et al., 2011), parsing (Foster et al., 2011), dialog modeling (Ritter et al., 2010) and summarization (Hutton and Kalita, 2010). These works adapt various parts of the natural language processing pipeline for social media text, and make use of a range of techniques (Preprocessing, New labeled data, New annotation schemes, Self training, Distributional features, Distance supervision) (Eisenstein, 2013).

Recently, Seddah et al. (2012) followed the second approach on French social media content and provided new labeled data and annotation schemes. They applied the MElt POS tagger (Denis and Sagot, 2012) embedded within text normalization and correction to noisy user generated texts and presented baseline POS tagging and statistical constituency parsing results.

3 Annotated Dataset

A set of 1,700 sentences (38k tokens) has been extracted from various types of French Web 2.0 user generated content (Facebook, Twitter, Video games and medical web forums) by Seddah et al. (2012). They selected these corpora through direct examination of various search queries and ranked the texts according to their distance from the French Treebank style, by measuring noisiness using the kullback-Leibler divergence between the distribution of trigrams of characters in given corpus and the distribution of trigrams of characters in the French Treebank reference. Some properties of this corpora are shown in Table 1.

They targeted the annotation scheme of the FTB-UC in order to annotate the French social media bank. The tagset includes 28 POS tags from FTB-UC and compound tags with additional categories specific to social media, including **HT** for Twitter hashtags and **META** for meta-textual tokens, such as

Twitter's "RT". Twitter at-mention as well as URLs and e-mail addresses have been tagged **NPP** which is the main difference with other works on on-line conversational texts. The inter-annotator agreement rate in this corpora range between 93.4% for **F**ACEBOOK data and 97.44% for **J**EUX**V**IDEOS.COM (Table 1) which indicates an almost perfect agreement on the corpus (Landis and Koch, 1977).

Corpus Name	**# sent.**	**# tokens**	Inter Annotator Agreement %
TWITTER	216	2465	95.40
FACEBOOK	452	4200	93.40
JEUX**V**IDEOS.COM	199	3058	97.44
DOCTISSIMO	771	10834	95.05

Table 1: Annotated datasets

4 Baseline

This section presents the performance of a state-of-the-art POS tagger for French, conducted by Seddah et al. (2012). They used FTB-UC as training, development and test data. First, they applied several correction processes in order to wrap the POS tagger to tag a sequence of tokens as close as possible to standard French and training corpus. Then, the MElt tagger has been used with a set of 15 language-independent rules, that aim at assigning the correct POS to tokens that belong to categories not found in training corpus (e.g., URLs, e-mail addresses, emoticons). The preliminary evaluation experiments with normalization and correction wrapper showed *84.72%* and *85.28%* token accuracy over annotated development and test set respectively.

5 New POS Tagger Development

Conversational style context and 140-character limitation in micro-blogs require users to express their thought or reply to others' messages within a short text. Therefore, without being ambiguous, some words are usually abbreviated with a special spelling. For example, *c t* usually means c'était (it was); *qil* denotes qu' il (that it/he).

Our tagger is based on sequence labeling models (CRF), enabling arbitrary local features to be integrated into a log-linear model. We employed three categories of feature templates to deal with syntactic variations on social media contents and alleviating the data sparseness problem.

5.1 Basic Feature Templates

The feature templates we use here are a superset of the largely language independent features used by (Ratnaparkhi, 1996; Toutanova and Manning, 2000; Toutanova et al., 2003). These features fall into two main categories. A first set of features tries to capture the *lexical form* of the word being tagged: it includes prefixes and suffixes (of at most 10 characters) from the current word, together with binary features based on the presence of special characters such as numbers, hyphens, and uppercase letters, within w_i. A second set of features directly models the context of the current word and tag: it includes the previous tag, surrounding word forms in a 5 tokens window. The detailed list of feature templates we used in this category is shown in Table 2.[1]

Context	
$w_i = X$,$i \in [-2,-1,0,1,2]$	& $t_0 = T$
$w_i w_j = XY$, $(i,j) \in \{(-1,0),(0,1),(-2,0),(0,2)\}$	& $t_0 = T$
$w_i w_j w_k = XYZ$, $(i,j,k) \in \{(-2,-1,0),(0,1,2),(-1,0,1)\}$	& $t_0 = T$
$w_i w_j w_k w_l w_m$=$XYZPQ$, (i,j,k,l,m) =$(-2,-1,0,1,2)$	& $t_0 = T$
t_{-1}	& $t_0 = T$
Lexical and Orthographic	
$f(w_i)$,$i \in [-1,0,1]$,$f \in F$	& $t_0 = T$
$m(w_i)$, $i \in [-1,0,1]$, $m \in M$	& $t_0 = T$

Table 2: Basic Feature Templates

[1] w_0 means the token at the current position while w_{-1} means the previous token.

The model generates the feature space by scanning each pair in the training data with the feature templates given in Table 2. For example, if we consider the following tweet from the training set, the generated features based on the first template can be seen in Table 3, in which the current word is "vous" (position 6) .

Sample tweet : "@Marie Je vais tener De vous produire la vidéo *-* "

word:	@Marie	Je	vais	tener	De	vous	produire	la	vidéo	*-*
Tag:	NPP	CLS-SUJ	V	VINF	P	CLO-A_OBJ	VINF	DET	NC	I
Position:	1	2	3	4	5	6	7	8	9	10

w_0=**vous**	$\&t_0$=**O**
w_{-1}=**De**	$\&t_0$=**O**
w_{-2}=**tener**	$\&t_0$=**O**
w_{+1}=**produire**	$\&t_0$=**O**
w_{+2}=**la**	$\&t_0$=**O**

Table 3: Generated features with template :
$w_i = X$,$i \in [-2, -1, 0, 1, 2]$ $\quad \&t_0 = T$

We defined two sets of operations, F and M. Each operation maps tokens to equivalence classes.
F is a set of regular expression rules that detect specific patterns on w_i and return binary values. The functions $f(w_i) \in F$ include the rules as detailed in the following list (List 1):

List 1: Set of regular expression rules (F)

- ▷ Return "$True$" if the w_i contains Punctuation marks otherwise return "$False$"
- ▷ Return "$True$" if the w_i is list of Punctuation marks otherwise return "$False$"
- ▷ Return "$True$" if the w_i contains digits otherwise return "$False$"
- ▷ Return "$True$" if the w_i number otherwise return "$False$"
- ▷ Return "$True$" if all letters of w_i are capitalized otherwise return "$False$" allNumber
- ▷ Return "$True$" if the w_i starts with capital letter otherwise return "$False$"
- ▷ Return "$True$" if the w_i has"URL" pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has "Email" pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has "Abbreviation" pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has "Arrow" pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has "Time " pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has "NumberWithCommas" pattern otherwise return "$False$"
- ▷ Return "$True$" if the w_i has symbol representing "RT:retweeting" form otherwise return "$False$"
- ▷ Return "$True$" if the w_i has symbol representing "At-Mention" form otherwise return "$False$"
- ▷ Return "$True$" if the w_i has symbol representing "hash-tagh" form otherwise return "$False$"

M is a set of orthographic transformations that maps a string to another string via a simple surface level transformation. The functions $m(w_i) \in M$ are given in List 2 :

List 2: Set of orthographic transformation (M)

- ▷ Return capitalized type of w_i ,These types are (allCap, shortCap, longCap, noCap, initCap, mixCap)
(e.g.,"Plus-tard" → "initCap" ,"RT" →"allCap,longCap")

- ▷ Return the *type* of w_i, obtained by replacing $[a-z]$ with x, $[A-Z]$ with X, and $[0-9]$ with 9
(e.g.,., "@DJRyan1der" → "@XXXxxx9xxx")

- ▷ Return a vector of Unicode matching of the string w_i
(e.g., "@DJRyan1der" → "$[64-68-74-82-121-97-110-49-100-101-114]$")

- ▷ Return the first n character of x (n-gram prefix), where $1 \leq n \leq 10$
- ▷ Return the last n character of x (n-gram suffix), where $1 \leq n \leq 10$

5.2 Word Clustering Feature Templates

To bridge the gap between high and low frequency words, we employed word clustering to acquire knowledge about paradigmatic lexical relations from large-scale texts. Our work is inspired by the suc-

cessful application of word clustering in supervised NLP models (Miller et al., 2004; Turian et al., 2010; Ritter et al., 2011; Owoputi et al., 2013).

Various clustering techniques have been proposed, some of which, for example, perform automatic word clustering optimizing a maximum likelihood criterion with iterative clustering algorithms. In this work, we focus on distributional word clustering, based on the assumption that the words that appear in similar contexts (especially surrounding words) tend to have similar meanings.

5.2.1 Brown Clustering

We used our unlabeled Twitter corpus (4M tweets) to improve our tagger performance. This corpus has been extracted in the framework of a French government funded ANR project called Imagiweb, whose goal is to develop tools to analyse the brand image of entities (persons or companies) on social media. More specifically, one of the focus of the project is to analyse the brand image of politicians on Twitter. Therefore, data about the two main candidates (F. Hollande and N. Sarkozy) in the last French presidential election in May 2012 have been crawled from Twitter, using Twitter API, from 6 months before to 6 months after the elections. Our unlabeled Twitter data is a sub-set of this corpus.

We obtained hierarchical word clusters via Brown Clustering (Brown et al., 1992) on a large set of unlabeled tweets. This algorithm generates a hard clustering, each word belongs to exactly one cluster. The input to the algorithm is a sequence of words $w_i, \ldots, w_n$. Initially, the algorithm starts with each word in its own cluster. As long as there are at least two clusters left, the algorithm merges the two clusters that maximize the resulting cluster quality. The quality is defined on the class-based bigram language model as follows, where C maps a word w to its class $C(w)$.

$$p(w_i|w_1, \ldots, w_{i-1}) = p(C(w_i)|C(w_{i-1}))p(w_i|C(w_i))$$

We ended up with 500 clusters (the optimal number of clusters according to the performance of the tagger among different number of clusters) with 222,788 word types by keeping the words appearing 10 or more times. Since Brown clustering creates hierarchical clusters in a binary tree, we used the feature template which maps the word w_i to the cluster at depths $2, 4, \ldots, 16$ containing w_i. If w_i was not seen while constructing the clusters and thus does not belong to any cluster we tried to find similar words by computing *Jaro-Winkler distance* (Philips, 1990; Winkler, 2006) and mapped the best match to the cluster depths. Nevertheless, if we couldn't find the best match (the threshold of the similarity score is 0.9), we mapped it to a special $NULL$ cluster. The detailed list of feature templates we used in this category is shown in Table 5.[2]

5.2.2 MKCLS Clustering

We also did some experiments, using another popular clustering method based on the exchange algorithm (Kneser and Ney, 1993). The objective function maximizes the likelihood $\prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1})$ of the training data given a partially class-based bigram model of the form as follows:

$$p(w_i|w_1, \ldots, w_{i-1}) \approx p(C(w_i)|w_{i-1})p(w_i|C(w_i))$$

We use the publicly available implementation MKCLS [3] to train this model on our French Twitter data (4M tweets). This algorithm provides us with 500 word clusters with 2,768,297 different words.

Word Cluster	
$c(w_i) = X$,$i \in [-2, -1, 0, 1, 2]$ and $c \in C$	& $t_0 = T$
$c(w_i)c(w_j) = XY$, $(i, j) \in \{(-1, 0), (0, 1)\}$ and $c \in C$	& $t_0 = T$
$c(w_i)C(w_j)c(w_k) = XYZ$, $(i, j, k) \in \{(-2, -1, 0), (0, 1, 2), (-1, 0, 1)\}$ and $c \in C$	& $t_0 = T$
$c(w_i)c(w_j)c(w_k)c(w_l)c(w_m)$=$XYZPQ$, (i, j, k, l, m) =$(-2, -1, 0, 1, 2)$ and $c \in C$	& $t_0 = T$

Table 5: Word Clustering Feature Templates

[2] $c(w_i) \in C$ map the word w_i to the clusters at depths $2, 4, \ldots, 16$

[3] `https://code.google.com/p/giza-pp/`

6 Experiments

For the implementation of discriminative sequential model, we chose the *Wapiti* [4] toolkit (Lavergne et al., 2010). Wapiti is a very fast toolkit for segmenting and labeling sequences with discriminative models. It is based on maxent models, maximum entropy Markov models and linear-chain CRF and proposes various optimization and regularization methods to improve both the computational complexity and the prediction performance of standard models. Wapiti has been ranked first on the sequence tagging task for more than a year on MLcomp[5] web site.

6.1 Training and parameter regularization

In the training of log-linear models, regularization is normally required to prevent the model from over fitting on the training data. The two most common regularization methods are called L1 and L2 regularization (Tsuruoka et al., 2009). *Wapiti* uses the elastic-net penalty of the form:

$$\rho_1 * |\theta|_1 + \frac{\rho_2}{2} * ||\theta||_2^2$$

and it is implemented with 3 different algorithms: *Orthant-Wise Limited-memory Quasi-Newton* (OWL-QN: L-BFGS), *Stochastic Gradient Descent* (SGD) and *Block Coordinate Descent*. We trained with *L-BFGS*, a classical Quasi-Newton optimization algorithm with limited memory which minimizes the regularized objective and uses elastic net regularization. Using even a very small L1 penalty excludes many irrelevant or highly noisy features. We carried out a grid search for the regularization values, assessing with F-measure and accuracy. We conducted a first order linear chain CRF model on the French corpora with classical setting (training set: 80%, development set: 10% and test set: 10%) for $L1 \in \{0, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$ and $L2 \in \{0, 0.0325, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$ (Owoputi et al., 2013). In any experiment, the result of the regularization values were close to each other, therefore we selected L1,L2=$(0.25, 0.5)$ achieving 80.4% and 90.6% F-measure and accuracy on the corpora respectively.

6.2 Performance

In order to assess how the results of our tagger based on the current limited corpora could be generalized to an independent data set, a set of 10-fold cross validation experiments has been performed. We investigated the effect of each feature template on the tagging. We used *"c: compact"* option in *Wapiti* which enables model compaction at the end of the training. This removes all inactive observations from the model, leading to a much smaller model when an L1-penalty is used.
Table 6 shows the result of each experiment, measured by token and sentence accuracy. It shows that word clustering is a very strong source of lexical knowledge and significantly increases the performance of our tagger.

Feature Templates	Token Accuracy %	Sentence Accuracy %
B	88.2	45.8
B+C1	90.8	49.9
B+C2	90.3	50.3
B+C1+C2	91.9	51.1

B: Basic Feature Templates
C1: Brown word-Clustering Feature Templates
C2: MKCLS word-Clustering Feature Templates

Table 6: Performance of new tagger based on CRF with different configurations

The CRF model with all set of features (B+C1+C2) is the best model with 91.9% and 51.1% token and sentence accuracy on 10-fold cross validation. All of these tagging accuracies are significantly above previous results on the French social bank (baseline).

[4] http://wapiti.limsi.fr/
[5] http://mlcomp.org/

7 Evaluation on English social media Content

In order to implement a tagger for English dedicated to social media content, we used the publicly available clusters data set (Owoputi et al., 2013) to build Brown clustering features. Moreover we performed the same process as in Section 5.2.2 in order to provide MKCLS clustering features with English Twitter data (1 million tweets obtained from [6]).

We applied our tagger with the best configuration to the annotated dataset provided by Ritter et al. (2011). This dataset contains 800 tweets that have been annotated with the Penn Treebank (PTB) tagset (Marcus et al., 1993). We trained and test our system with 10-fold cross validation. Table 7 shows our tagger performance compared to other state-of-art taggers on this data set.

Tagger	Accuracy%
Our new tagger, CRF with B+C1+C2 configuration	90.1
Ritter et al. (Ritter et al., 2011), CRF tagger	88.3
Owoputi et al. (Owoputi et al., 2013), MEMM tagger	90± 0.5

Table 7: Evaluation on Twitter data with PTB tags

In addition, we evaluated the tagger performance on another English social media data: NPS chat ("Chat with PTB tags" (Forsythand and Martell, 2007)). Due to the large number of tokens (50 K), we trained and tested our tagger with a 5-fold cross validation setup. Our new tagger performance as well as the other taggers results are given in Table 8.

Tagger	Accuracy%
Our new tagger, CRF with B+C1+C2 configuration	92.7
Forsythand and Martell (Forsythand and Martell, 2007), HMM tagger	90.8
Owoputi et al. (Owoputi et al., 2013), MEMM tagger	93.4± 0.3

Table 8: Evaluation on Chat data with PTB tags

8 Conclusion and Future Work

In this paper, we have presented an innovative work on POS tagging for French social media noisy input. Because of the specific phenomena encountered in such data and also because of the lack of large training corpus, we proposed a discriminative sequence labeling model (CRF) enhanced with several type of features. After experimenting different configurations of features, we achieved 91.9% token accuracy on target corpus. Moreover, experiments on English social media contents show that our model obtains further improvement over previous works on these data and could be reproduced for other languages. In the future, we plan to pursue this work in two main directions: (a) Integrate the new tagger with a robust syntactic parser and investigate its impact on dependency parsing applied to social media and (b) evaluate the impact of POS tagging on opinion mining on micro-blogs, since this parser is the core component of an opinion mining system applied in different social-media analytics projects.

References

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

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, pages 138–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, 46:721–736.

[6] http://illocutioninc.com/site/products-data.html

Jacob Eisenstein. 2013. What to do about bad language on the internet. *In proc. of NAACL.*

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88.

Jenny R. Finkel, Trond Grenager, and Manning Christopher. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of ACL*, pages 363–370.

Eric N. Forsythand and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 19–26, Washington, DC, USA. IEEE Computer Society.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Joseph Van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of web 2.0. *In Proceedings of IJCNLP*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 2:42–47.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: makn sens a #twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1:365–378.

Beaux Sharifi Mark-Anthony Hutton and Jugal Kalita. 2010. Summarizing microblogs automatically. *In Proceedings of NAACL.*

Reinhard Kneser and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modeling. In *In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Diana Maynard, Kalina Bontcheva, and Dominic Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at International Conference on Language Resources and Evaluation, LREC 2012, 26 May 2012, Istanbul, Turkey*.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT*, pages 337–342.

O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N.A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. *In Proceedings of NAACL.*

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language*, 7:12.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. *In Proceedings of NAACL.*

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. *ACL.*

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *COLING 2012 - 24th International Conference on Computational Linguistics*, Mumbai, India, Dec. Kay, Martin and Boitet, Christian.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *In EMNLP/VLC 2000*, pages 63–70.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *IN PROCEEDINGS OF HLT-NAACL*, pages 252–259.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 477–485, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*.

William E Winkler. 2006. Overview of record linkage and current research directions. Technical report, BUREAU OF THE CENSUS.

Liu Xiaohua, Zhang Shaodian, Wei Furu, and Zhou Ming. 2011. Recognizing named entities in tweets. *In Proceedings of ACL*.