Domain Adaptation
for Named Entity Recognition Using CRFs

Tian Tian[*,†], **Marco Dinarelli**[*], **Isabelle Tellier**[*], **Pedro Dias Cardoso**[†]
[*]LaTTiCe (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle - Paris 3
PSL Research University, USPC (Université Sorbonne Paris Cité)
1 Maurice Arnoux 92120 Montrouge FRANCE
isabelle.tellier@univ-paris3.fr, marco.dinarelli@ens.fr
[†] Synthesio
8-10 rue Villedo 75001 Paris FRANCE
ttian, pedro@synthesio.com

Abstract

In this paper we explain how we created a labelled corpus in English for a Named Entity Recognition (NER) task from multi-source and multi-domain data, for an industrial partner. We explain the specificities of this corpus with examples and describe some baseline experiments. We present some results of domain adaptation on this corpus using a labelled Twitter corpus (Ritter et al., 2011). We tested a semi-supervised method from (Garcia-Fernandez et al., 2014) combined with a supervised domain adaptation approach proposed in (Raymond and Fayolle, 2010) for machine learning experiments with CRFs (Conditional Random Fields). We use the same technique to improve the NER results on the Twitter corpus (Ritter et al., 2011). Our contributions thus consist in an industrial corpus creation and NER performance improvements.

Keywords: Domain Adaptation, Social Media, CRFs, Machine Learning

1. Introduction

Social media (forums, Facebook, Twitter, etc) are now becoming the major use of Internet. Within these platforms, more and more topics are being discussed everyday. The automatic analysis of these massive data is a challenge, as the texts produced in these contexts differ from previously available texts. Some works have already been dedicated to traditional Natural Language Processing (NLP) tasks such as part-of-speech tagging (Gimpel et al., 2011), and even complete parsing (Foster et al., 2011) to social media data like texts from Twitter. We focus here on the task of Named Entity Recognition (NER) with both long text and short text (like tweets). NER is a traditional NLP task (see for example the CoNLL shared task 2003 (Tjong Kim Sang and De Meulder, 2003), with texts from Reuters) which has also already been addressed for tweets (Ritter et al., 2011).

In this paper, we first describe, in Section 2, a special multi-domain and multi-source NER task in English from raw data (most of them from social media) for an industrial partner. In Section 3, we explain how we created an annotated reference corpus from four different domains and two different sources (mostly forums and Twitter) to evaluate the NER task. The Section 4 contains the description of our baseline, a Conditional Random Fields (CRFs) model trained with a Twitter corpus. We then conducted some domain adaptation experiments, whose results are provided in Section 5. In these experiments, we used an iterative training method with unlabelled data from one domain. We finally suggest ways to further improve our results, especially the recall value.

2. Task

This work aims at helping Synthesio[1] to better analyse its data automatically. Synthesio is a social listening platform, providing scalable monitoring and analytic solutions to hundreds of brands and agencies around the world. Its clients use this service to cut through social noise, find the conversations that matter, measure their online reputation, manage their consumer relationships and boost the Return On Investment (ROI) of their social activities.

From the mass of data from social media, including Twitter, forums and Facebook, Synthesio should first find those discussing about one of their customers (a client company or a brand, for example), and then try to analyse the distribution of opinions presented in these data.

The first step toward this processing is a Named Entity Recognition (NER) task, which aims to find the brand/company/person/product a text talks about. Some brands like “Boss” (clothes), “President” (cheese) are also common nouns, so the task is not easy and often includes disambiguition.

For the needs of Synthesio, we have defined 9 types of entities, shown in Table 1.

Company	Company name
Person	Person name
Geo-loc	Location, country or city name
Facility	Organization name
Product	Product name
Media	Journal, music artist
Sportsteam	Sports team name
Job-title	Job names like director, PDG
Other	Holidays, events, etc

Table 1: Synthesio Named Entity Definition.

A single token (or a sequence of tokens) can in general belong to different classes. For example,

[1] www.synthesio.com

> My McDonald's was still hot when it was served and it tasted delicious.

and another text :

> The room was too hot when we ate at McDonald's yesterday.

Although both texts talk about "McDonald's", the first one is about a food product, and the second one about a place. Furthermore, one has positive opinion while the other has a negative one, still using very similar words. The context is important to distinguish these cases.

3. Corpus

For each client, Synthesio collects as many texts as possible from official web sites, discussion forums, Facebook pages and Twitter posts through the Internet. These texts are raw data: some of them appear in double (because of retweets), some are not analysable (tweets with only hashtags #), etc. There is no annotated data available for this NER task.

Moreover, Synthesio's clients vary from cosmetic and children's toys to automobile and fastfoods. Each domain has its own vocabulary and specific expressions. That makes this NER task difficult.

In order to evaluate an automatic system for this NER task with data from Synthesio, we first need a reference corpus. So we chose five different clients from four domains where Synthesio has the most clients: Deezer (music), Dunkin Donuts (food and coffee), Mattel (toy), Land Rover and Nissan (automobile). For each of these five clients, we distinguished resources from journals and forums (long text data) and from Twitter, instagram, etc. (short text data), and we extracted 50 texts from each resource for each client by a simple request to Synthesio search engine. This extraction is random, but somehow certain texts are really similar or have exactly the same content, but with different internal ids. In the case of Twitter, retweets have exactly the same contents ("RT : A" where A is the repeated text).

Synthesio keeps these multicopies in its database for tracing tweets history, but these double texts are not useful to create a reference corpus. So we use a score to measure the similarity of every pair of candidate texts (i, j) to be added to the reference corpus. Both texts are kept when:

$$\frac{V_i \cap V_j}{V_i \cup V_j} < 0.6 \tag{1}$$

where V_i and V_j are the vocabularies of i and j, respectively. The formula is a Dice similarity applied to a "bag of words" representation of i and j. If it is not true, both texts are considered as too similar and only one of them is kept in the reference corpus. Table 2 shows some statistics about this Synthesio reference corpus after this filter is applied.

These Synthesio raw data are partly made of texts from Twitter (short text data). The other part (long text data) from Facebook contains longer texts which are also mostly not written in grammatically well-formed English. The Named Entity types we try to extract are similar to those of the Ritter NER corpus (Ritter et al., 2011) from Twitter. That's why we chose the Ritter NER corpus as starting point to pre-label our reference corpus with named entities. The Ritter NER corpus contains 2394 sequences, that is 46469 tokens. First, we modified the Ritter corpus entity set by adding the "job-title" entity and merging "tv show", "movie" and "music artist" into the entity "media". The number of occurrences of each named entity in the 2 resulting corpora is shown in Table 3.

long text	Deezer	Dunkin	Land	Mattel	Nissan
sequences	146	208	183	174	123
tokens	2314	3116	3836	2958	2166

short text	Deezer	Dunkin	Land	Mattel	Nissan
sequences	52	50	59	57	74
tokens	854	827	1048	1123	1127

Table 2: Synthesio Reference Corpus.

	Ritter	Synthesio
Sequences	2194	1126
Tokens	48k	19k
Company	186	496
Product	102	484
Media	126	41
Job-title	87	18
Geo-loc	291	66
Person	472	83
Facility	107	11
Sportsteam	55	6
Other	246	13
total	1672	1812

Table 3: Modified Ritter Named Entity Corpus and Synthesio reference corpus.

From the table 3 we can see that the entity distribution is very different in the 2 corpora. The size of Ritter Corpus is twice larger than Synthesio reference corpus. As we can see the number of `Company` and `Product` entities in the Synthesio corpus is almost four times more than in the Ritter Corpus. On the other hand, the number of `Person`, `Geo-loc` and `job-title` entities in the Synthesio corpus is only about one fifth of the same entities in the Ritter Corpus.

We have trained a CRF model using this corpus, using a simple unigram version of (Lavergne et al., 2010) CRFs template. We used this CRF model to label our reference corpus. Afterwards, an annotator manually corrected the annotations. Entities with # or @ were not annotated in the Ritter corpus. But Synthesio required this annotation when the corresponding entity was relevant.

Although we spent a lot of time and efforts on this reference corpus, it is still not large enough to train an effective model for one domain. For this reason we decided to apply domain adaptation approaches. These allow to exploit unlabelled data from a given target domain, and can be much more abundant then labelled data.

4. Related Work

Domain adaptation has been discussed for many machine learning techniques in many NLP tasks: pos-tagging and chunking in (Xiao and Guo, 2015), named entity recognition in (Guo et al., 2009) and (Yu and Jiang, 2015), opinion mining in (Garcia-Fernandez et al., 2014) and (Blitzer et al., 2007), relation extraction in (Nguyen et al., 2015) and spam detection in (Yu and Jiang, 2015).

The underlying ideas are similar. Like for machine learning, there are three general approaches: supervised, unsupervised and semi-supervised.

In supervised domain adaptation, (Daumé et al., 2010) augment the number of features of their data representation to distinguish regularities of source and target corpora. They multiply every single feature by 3: source domain features, target domain features and general features from source and target corpora. Similarly, (Raymond and Fayolle, 2010) try to use few features from the source corpus and to complete with more features from the target one in order to make the latter to dominate on source domain features. (Jiang and Zhai, 2007) follow the same idea but uses distributional representations. (Arnold et al., 2008) propose a hierarchical structure. (Blitzer et al., 2006) and (Xiao and Guo, 2015) use a distribution representation for tokens. These methods aim to use different weights for features from source and target corpora.

In unsupervised domain adaptation, (Freitag, 2004) use clustering to group words into sets. (Yu and Jiang, 2015) develop a similar approach with unlabelled data to add them into the training set.

As for semi-supervised domain adaptation, (Nguyen et al., 2015) use lexical semantic representations to enrich the training data. (Garcia-Fernandez et al., 2014) try an iterative training procedure to add predictions concerning unlabelled data into a labelled training set.

5. Methods and Results

5.1. Baseline

As mentioned before, we automatically annotated the Synthesio reference corpus with a CRF model trained on the Ritter corpus and corrected the annotation manually. We found out that the Synthesio data are very different from the Ritter corpus. The Table 4 shows the baseline evaluation in terms of F1 mesure micro-average of the model trained on the Ritter corpus and tested on the 5 domains of the Synthesio data.

Long text	Deezer	Dunkin	Land	Mattel	Nissan
Precision	0.56	0.09	0.31	0.18	0.02
Recall	0.14	0.06	0.05	0.07	0.01
F1-measure	0.22	0.07	0.08	0.08	0.01
Short text	Deezer	Dunkin	Land	Mattel	Nissan
Precision	0.74	0.39	0.56	0.19	0.06
Recall	0.09	0.08	0.05	0.02	0.02
F1-measure	0.16	0.13	0.09	0.03	0.03

Table 4: Baseline evaluated with reference corpus from different sources.

We can see that the model trained on the Ritter corpus doesn't perform well on Synthesio data, compared to cross validation results on the Ritter corpus in (Ritter et al., 2011). This clearly shows that these data come from a different domain.

Meanwhile, all Synthesio data are also different because they are from different domains (food and coffee, automobile, etc). Moreover none of the annotated part of these corpora is big enough to build an effective model.

In contrast, Synthesio has large amount of unlabelled texts in its database. We thus tried to exploit these unlabelled data to improve the NER task results, using the model trained on the Ritter corpus, via domain adaptation.

5.2. Iterative training

We first extracted all texts in the `Deezer` domain published during one day (2015 October 5th) in forums, blogs etc., what we call *long texts* (in contrast to texts coming from tweets, which are shorter). We then filtered repeated text sequences and similar text sequences using equation1. So we obtained an unlabelled long text `Deezer` data with $1M$ sequences, that is more than $41M$ tokens. These data are quite noisy, with some sequences really different from well-formed English texts.

Our iterative training procedure follows (Garcia-Fernandez et al., 2014). The idea is to annotate unlabelled data with an initial model (here the CRF model trained on the Ritter corpus). We then pick up all annotated sequences for which the model has a confidence score higher than a given threshold. We add these sequences into the initial data to train a new model. In this step, we keep the same features for the Ritter corpus and the predicted Synthesio data. This process repeats until no more sequence passes the threshold. This procedure was originally applied to sentiment classification. In the NER task, where most of the labels are "O" ("Outside" an entity in a BIO annotation), most sequences which pass a high threshold (for example 0.9) were predicted with only "O" labels. In order to add only meaningful sequences to the training data, we add only sequences which pass the threshold and contain at least one named entity.

We chose first a high threshold of 0.8 and there were 1608 sequences with more than one entity. Then for each Synthesio domain, kept as test domain out of five domains, we trained a mixed model with the Ritter corpus, plus these 1608 predicted sequences and the other eight corpora in the other domains (4 of *short texts* and 4 of *long texts*). We can evaluate thus on all Synthesio reference corpora.

long text	Deezer	Dunkin	Land	Mattel	Nissan
Precision	0.68	0.5	0.23	0.55	0.58
Recall	0.13	0.24	0.08	0.28	0.09
F1-measure	0.22	0.32	0.12	0.37	0.16
short text	Deezer	Dunkin	Land	Mattel	Nissan
Precision	0.5	0.6	0.05	0.49	0.3
Recall	0.08	0.23	0.06	0.19	0.09
F1-measure	0.14	0.33	0.05	0.27	0.14

Table 5: Evaluation of a mixed model trained with the Ritter corpus and predicted sequences with a confidence score more higher than 0.8.

The table 5 shows the results of this procedure with only a first iteration. As we can see the model performs far better than the baseline 4 for Dunkin Donuts, Mattel and Nissan, but not for Deezer. Among the 1608 sequences annotated, even the word "Deezer" was not labelled as an entity. Since we have much more selected data than the original corpus

Ritter, we also tried to filter annotated text sequences with a threshold of 0.9. The table 6 shows the number of selected sequences for each domain after first prediction.

Domain name	Accepted	Entities
long Nissan	9	9
short Nissan	6	6
long Mattel	1010	1056
short Mattel	19	24
long LandRover	2846	3057
short LandRover	47	48
long DunkinDonuts	22550	24156
short DunkinDonuts	102008	103176
long Deezer	1054	1105
short Deezer	32469	33166

Table 6: Number of selected sequences in each domain

From these selected sequences, we can remarque some examples:

> Argyle fan in North Yorkshire.

Here our model annotated the "Geo-loc" entity "North Yorkshire" even when this phrase is absent in our training data, which means that model reaches a certain generalisation capability.

> but people are saying that she's a witch doctor.

Here is an example of boundary error. This "job-title" entity should be the whole phrase "witch doctor" but our model extracted only "doctor".

> Nissan Note Nismo Coming This Fall in Japan.

In this example, neither company "Nissan" nor product "Note Nismo" is extracted as entity, but "Japan" is annotated correctly as "Geo-loc" entity.

As shown in table 6, the unsupervised Synthesio data is not homogene. Some domains have more selected data than others. We leave as future work using these selected data (perhaps with a higher confidence probability) to create a more suitable training dataset, and maybe to use an entity list as CRF feature to improve the recall.

5.3. Iterative training with reduced features

Since Synthesio data are different from those of the Ritter corpus, we consider them like two different domains. Following annotation adaptation approach in (Raymond and Fayolle, 2010), we consider the Synthesio data as the target domain, the Ritter corpus as the source domain. We thus use the complete set of features for Synthesio data and only token values and pos tags (from Synthesio pos-tagger which tags 17 grammatical categories) for the Ritter Corpus. This affects the importance given by the CRF model to features extracted from the two corpora, giving more weight to those extracted from the target domain. Here, the best model is the one which takes into account only a window of size 3 (that is tokens in positions -1, 0 and 1).

The table 7 shows our results for the long-text and short-text Deezer corpora with a model trained with Ritter plus the Synthesio corpus with the full set of features (baseline templates). Here the Synthesio corpus contains the 8 reference corpora provided by Synthesio and the 1608 sequences obtained in the previous experiment.

Model	Corpus	Precision	Recall	F1
full features	long text	0.83	0.08	0.15
	short text	0.42	0.03	0.06
token1+pos5	long text	0.63	0.06	0.11
	short text	0.52	0.03	0.06
token3+pos3	long text	0.67	0.06	0.11
	short text	0.41	0.04	0.07

Table 7: Evaluation of model with Ritter on reduced features with Deezer.

We can see that compared to the "full features" model, the model with reduced features gives a slightly better result on Twitter data (short-text corpus).

6. Conclusion and future work

In this paper, we explained the multi-domain and multi-source NER task for Synthesio. We created a reference corpus from two sorts of sources and four domains. Since the CRF model trained on the Twitter text (Ritter corpus) performs poorly on this reference corpus, we showed a way to use large unlabelled data to improve NER results. This is done by iteratively incrementing the training data with automatically annotated sequences having a prediction confidence higher than a given threshold. We also tried to combine this domain adaptation technique with different features selections from the source data and target domain.

As future work, we'll try to improve the recall using some entity list as features in our CRFs models, and use Synthesio Twitter data to improve the Ritter NER results.

7. Bibliographical References

Arnold, A., Nallapati, R., and Cohen, W. W. (2008). Exploiting feature hierarchy for transfer learning in named entity recognition. In *Proceedings of ACL-08: HLT*, pages 245–253, Columbus, Ohio, June. Association for Computational Linguistics.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 187–205.

Daumé, III, H., Kumar, A., and Saha, A. (2010). Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., Van Genabith, J., et al. (2011). # hardtoparse: Pos tagging and parsing the twitterverse. In *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25.

Freitag, D. (2004). Trained named entity recognition using distributional clusters. In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 262–269, Barcelona, Spain, July. Association for Computational Linguistics.

Garcia-Fernandez, A., Ferret, O., and Dinarelli, M. (2014). Evaluation of different strategies for domain adaptation in opinion mining. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guo, H., Zhu, H., Guo, Z., Zhang, X., Wu, X., and Su, Z. (2009). Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 281–289, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June. Association for Computational Linguistics.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 504–513, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nguyen, T. H., Plank, B., and Grishman, R. (2015). Semantic representations for domain adaptation: A case study on the tree kernel-based method for relation extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 635–644, Beijing, China, July. Association for Computational Linguistics.

Raymond, C. and Fayolle, J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Conférence Traitement automatique des langues naturelles, TALN'10*, Montréal, Québec, Canada, July. ATALA.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao, M. and Guo, Y. (2015). Learning hidden markov models with distributed state representations for domain adaptation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–529, Beijing, China, July. Association for Computational Linguistics.

Yu, J. and Jiang, J. (2015). A hassle-free unsupervised domain adaptation method using instance similarity features. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 168–173, Beijing, China, July. Association for Computational Linguistics.