Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 44 (2017) 61–83

www.elsevier.com/locate/csl

Generalisation in named entity recognition: A quantitative analysis

Isabelle Augenstein*, Leon Derczynski, Kalina Bontcheva

University of Sheffield, Sheffield, S14DP, UK

Received 5 January 2016; received in revised form 22 August 2016; accepted 1 January 2017
Available online 14 February 2017

Abstract

Named Entity Recognition (NER) is a key NLP task, which is all the more challenging on Web and user-generated content with their diverse and continuously changing language. This paper aims to quantify how this diversity impacts state-of-the-art NER methods, by measuring named entity (NE) and context variability, feature sparsity, and their effects on precision and recall. In particular, our findings indicate that NER approaches struggle to generalise in diverse genres with limited training data. Unseen NEs, in particular, play an important role, which have a higher incidence in diverse genres such as social media than in more regular genres such as newswire. Coupled with a higher incidence of unseen features more generally and the lack of large training corpora, this leads to significantly lower $F1$ scores for diverse genres as compared to more regular ones. We also find that leading systems rely heavily on surface forms found in training data, having problems generalising beyond these, and offer explanations for this observation.

Keywords: Natural language processing; Information extraction; Named entity recognition; Generalisation; Entity drift; Social media; Quantitative study

1. Introduction

Named entity recognition and classification (*NERC*, short *NER*), the task of recognising and assigning a class to mentions of proper names (named entities, *NEs*) in text, has attracted many years of research (Nadeau and Sekine, 2007; Ratinov and Roth, 2009), analyses (Palmer and Day, 1997), starting from the first MUC challenge in 1995 (Grishman and Sundheim, 1995). Recognising entities is key to many applications, including text summarisation (Schiffman et al., 2002), search (Toda and Kataoka, 2005), the semantic web (Maynard et al., 2016), topic modelling (Newman et al., 2006), and machine translation (Al-Onaizan and Knight, 2002; Steinberger et al., 2011).

As NER is being applied to increasingly diverse and challenging text genres (Derczynski et al., 2013, Fromreide et al., 2014, Whitelaw et al., 2008), this has lead to a noisier, sparser feature space, which in turn requires regularisation (Cherry and Guo, 2015) and the avoidance of overfitting. This has been the case even for large corpora all of the same genre and with the same entity classification scheme, such as ACE (Mooney and Bunescu, 2005). Recall, in particular, has been a persistent problem, as named entities often seem to have unusual surface forms, e.g. unusual

* Corresponding author.
E-mail addresses: i.augenstein@ucl.ac.uk, isabelle.augenstein@gmail.com (I. Augenstein).

http://dx.doi.org/10.1016/j.csl.2017.01.012

character sequences for the given language (e.g. *Szeged* in an English-language document) or words that individually are typically not NEs, unless they are combined together (e.g. *the White House*).

Indeed, the move from ACE and MUC to broader kinds of corpora has presented existing NER systems and resources with a great deal of difficulty (Maynard et al., 2003), which some researchers have tried to address through domain adaptation, specifically with entity recognition in mind (Daumé, 2007; Wu et al., 2009; Guo et al., 2009; Chiticariu et al., 2010; Augenstein, 2014). However, more recent performance comparisons of NER methods over different corpora showed that older tools tend to simply fail to adapt, even when given a fair amount of in-domain data and resources (Ritter et al., 2011; Derczynski et al., 2015). Simultaneously, the value of NER in non-newswire data (Ritter et al., 2011; Liu et al., 2011; Plank et al., 2014; Rowe et al., 2015; Baldwin et al., 2015) has rocketed: for example, social media now provides us with a sample of all human discourse, unmolested by editors, publishing guidelines and the like, and all in digital format − leading to whole new fields of research opening in computational social science (Hovy et al., 2015; Plank and Hovy, 2015; Preoţiuc-Pietro et al., 2015).

The prevailing assumption has been that this lower NER performance is due to domain differences arising from using newswire (NW) as training data, as well as from the irregular, noisy nature of new media (e.g. Ritter et al., 2011). Existing studies (Derczynski et al., 2015) further suggest that named entity diversity, discrepancy between named entities in the training set and the test set (*entity drift* over time in particular), and diverse context, are the likely reasons behind the significantly lower NER performance on social media corpora, as compared to newswire.

No prior studies, however, have investigated these hypotheses quantitatively. For example, it is not yet established whether this performance drop is really due to a higher proportion of unseen NEs in the social media, or is it instead due to NEs being situated in different kinds of linguistic context.

Accordingly, the contributions of this paper lie in investigating the following open research questions:

RQ1 How does NERC performance differ for corpora between different NER approaches?
RQ2 How does NERC performance differ for corpora over different text types/genres?
RQ3 What is the impact of NE diversity on system performance?
RQ4 What is the relationship between Out-of-Vocabulary (OOV) features (unseen features), OOV entities (unseen NEs) and performance?
RQ5 How well do NERC methods perform out-of-domain and what impact do unseen NEs (i.e. those which appear in the test set, but not the training set) have on out-of-domain performance?

In particular, the paper carries out a comparative analysis of the performance of several different approaches to statistical NER over multiple text genres, with varying NE and lexical diversity. In line with prior analyses of NER performance (Palmer and Day, 1997; Derczynski et al., 2015), we carry out corpus analysis and introduce briefly the NER methods used for experimentation. Unlike prior efforts, however, our main objectives are to uncover the impact of NE diversity and context diversity on performance (measured primarily by $F1$ score), and also to study the relationship between OOV NEs and features and $F1$. See Section 3 for details.

To ensure representativeness and comprehensiveness, our experimental findings are based on key benchmark NER corpora spanning multiple genres, time periods, and corpus annotation methodologies and guidelines. As detailed in Section 2.1, the corpora studied are OntoNotes (Hovy et al., 2006), ACE (Walker et al., 2006), MUC 7 (Chinchor, 1998), the Ritter NER corpus (Ritter et al., 2011), the MSM 2013 corpus (Rowe et al., 2013), and the UMBC Twitter corpus (Finin et al., 2010). To eliminate potential bias from the choice of statistical NER approach, experiments are carried out with three differently-principled NER approaches, namely Stanford NER (Finkel et al., 2005), SENNA (Collobert et al., 2011) and CRFSuite (Okazaki, 2007) (see Section 2.2 for details).

2. Datasets and methods

2.1. Datasets

Since the goal of this study is to compare NER performance on corpora from diverse domains and genres, seven benchmark NER corpora are included, spanning newswire, broadcast conversation, Web content, and social media (see Table 1 for details). These datasets were chosen such that they have been annotated with the same or very similar entity classes, in particular, names of people, locations, and organisations. Thus corpora including only domain-specific