Instituto Nacional de Ciência e Tecnologia para a Web

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPEMIG Fundação de Amparo à Pesquisa do Estado de Minas Gerais

LBD UFMG

A Lightweight Filter-Stream Approach for Named Entity Recognition on Twitter Data

Diego Marinho de Oliveira, Alberto H. F. Laender, Adriano Veloso, Altigran S. da Silva

Abstract: *Microblog platforms such as Twitter are being increasingly adopted by Web users, yielding an important source of data for web search and mining applications. Tasks such as Named Entity Recognition are at the core of many of these applications, but the effectiveness of the available recognition tools is seriously compromised when applied to Twitter data, since messages are terse (limited to 140 characters), poorly worded (e.g., containing misspellings, short forms and slangs) and posted in many different languages. Also, Twitter follows a streaming paradigm, imposing that entities must be recognized in real-time. In view of these challenges and the inappropriateness of existing tools, we propose a novel approach for Named Entity Recognition on Twitter data called FS-NER (Filter-Stream Named Entity Recognition). Through a systematic evaluation we show that, our proposed approach performs 3% better than a CRF-based baseline, besides being orders of magnitude faster and much more practical.*

Introduction

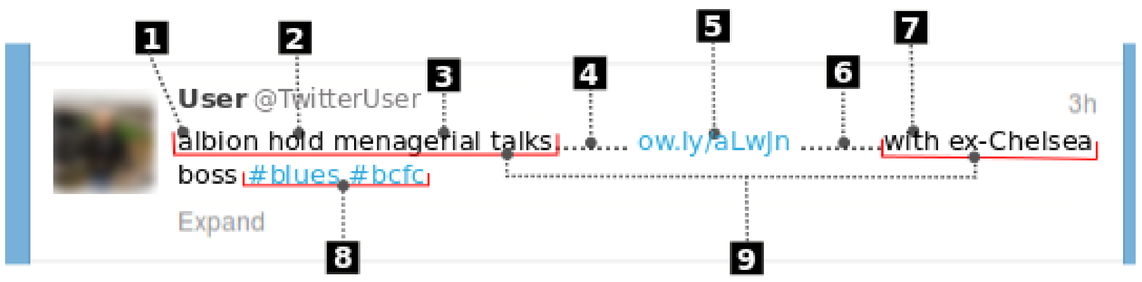
Microblogging activity is reshaping the way people communicate yielding a unique source of data for web search and mining. Applications that uses this source usually require identifying free-text references to named entities such as people, organizations, places, companies, and others [3] - a task commonly known as **Named Entity Recognition**.

Dominant approaches for Named Entity Recognition (NER) are either based on linguistic grammar-based techniques or on statistical models. Both approaches have demonstrated to be successful when applied to data obtained from typical Web documents, but they are ill suited when it comes to Twitter data [1]-[2].

In this work we propose a novel NER approach, called **FS-NER** (**Filter Stream Named Entity Recognition**). Essentially, the NER process in FS-NER is viewed as a coarse grain Twitter message flow controlled by a series of components, referred to as filters. A filter receives a Twitter message coming on the stream, performs specific processing in this message and returns information about possible entities in the message. Through our results, we show that, despite the simplicity of the filters used, the proposed approach is still able to outperform the baselines with improvements of 3% on average, while being orders of magnitude faster and thus more appropriate to the data streaming paradigm followed by Twitter.

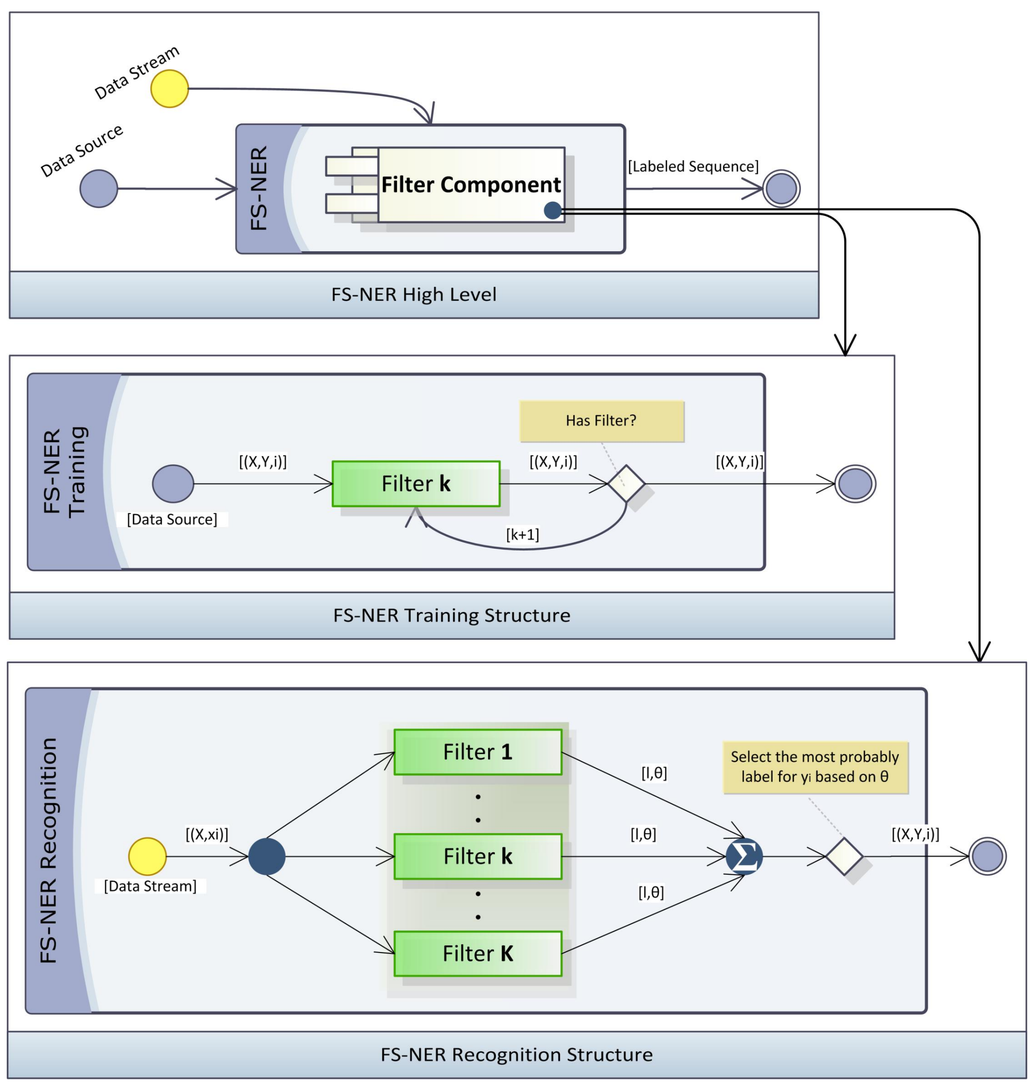
Challenges

- Large volume of data
- Lack of formalism



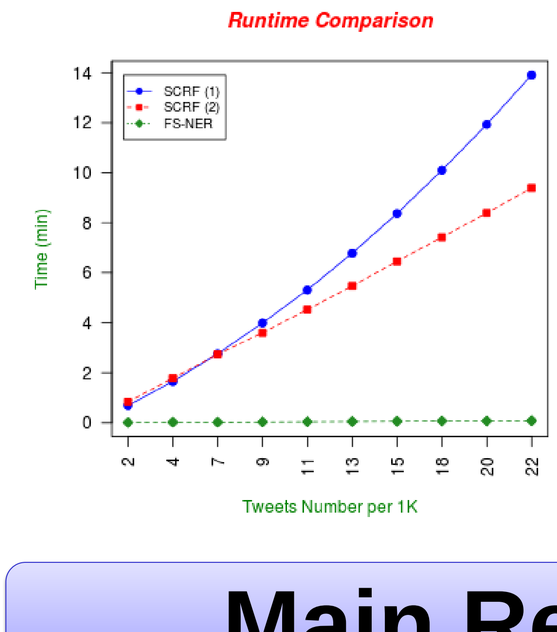
- Language dependece
- Real-time nature
- Lack of contextualized data
- Data stream orientation

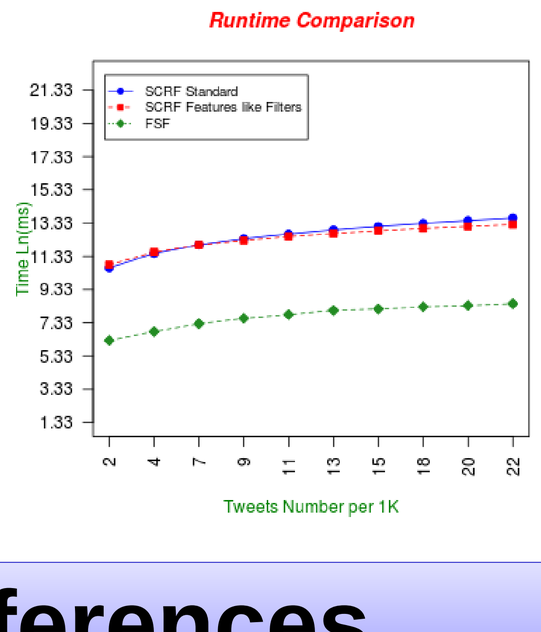
Approach



Experiments

Entity Type	RMCE	FS-NER	SCRF(2)	Diff.	t	p-value
Player	-	0.76±0.07	0.74±0.06	0.02	1.76	0.15
Venue	-	0.79±0.06	0.75±0.04	0.04	2.28	0.09
Team	-	0.86±0.01	0.86±0.01	0.00	-0.22	0.85
Company	0.58±0.07	0.49±0.10	0.50±0.10	-0.01	-1.02	0.36
Geo-loc	0.73±0.05	0.45±0.07	0.38±0.11	0.07	2.53	0.06
Person	0.78±0.04	0.57±0.02	0.46±0.04	0.11	5.39	0.01
Org	-	0.75±0.01	0.75±0.01	0.00	0.95	0.40
AVG	0.69	0.67	0.63	0.03	-	-
STDDEV	0.10	0.16	0.18	0.04	-	-





Main References

[1] *Alan Ritter, Sam Clark, Mausam, and Oren Etzioni.* **Named Entity Recognition in Tweets: An Experimental Study**. *In Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media, 2011.*

[2] *Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou.* **Recognizing Named Entities in Tweets**. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 359–367, 2011.*

[3] *David Nadeau and Satoshi Sekine.* **A Survey of Named Entity Recognition and Classification**. *Linguisticae Investigationes, 30(1):3–26, January 2007.*

Contato: **dmoliveira@ufmg.br**
http://dcc.ufmg.br/~dmoliveira