Building a treebank of noisy user-generated content: The French Social Media Bank

Djamé Seddah Benoît Sagot Marie Candito
Virginie Mouilleron Vanessa Combet
Alpage, Inria Paris-Rocquencourt & Université Paris Diderot, France

Abstract

We introduce the French Social Media Bank, the first user-generated content treebank for French. Its first release contains 1,700 sentences from various Web 2.0 and social media sources (FACEBOOK, TWITTER, web forums), including data specifically chosen for their high noisiness.

1 Introduction

New forms of electronic communication have emerged in the last few years, namely social media and Web 2.0 communication media, both synchronous (e.g., micro-blogging) or asynchronous (e.g., forums). These new user-generated contents often depart, sometimes heavily, from canonical language. This prevents an accurate processing of such data by current state-of-art NLP tools (Foster et al. [9], Gimpel et al. [11]). The main difficulties, highlighted by Foster [8], range from surface differences (intended or accidental non-standard typography) to lexical idiosyncrasies (genuine unknown words, sloppy spelling) and specific syntactic structures absent from well-edited data (imperatives, direct speech, slang, etc.).

The still difficult handling of those phenomena pleads for a better linguistic modeling and analysis of user-generated content. We introduce the French Social Media Bank, a freely available treebank containing 1700 manually annotated sentences. It constitutes the first resource covering the variety of French social medias, and the first data set we are aware of for FACEBOOK data.

2 Corpus

The French web 2.0 covers a wide range of practices. We decided to focus on microblogging (FACEBOOK and TWITTER) and on two types of web forums: one large-audience health forum, DOCTISSIMO (forum.doctissimo.fr) and one specialized on video games JEUXVIDEOS.COM (www.jeuxvideo.com). For each source but the latter, we gathered both lightly edited data and *noisier* data, using handcrafted search queries. Lightly edited data were retrieved based on source-specific news topics. The noisiest texts, intended to serve as a stress test for

	# sent.	# tokens	avg. lgth	std dev.	noisiness score
DOCTISSIMO	771	10834	14.05	10.28	0.37
high noisiness subcorpora	36	640	17.78	17.63	1.29
other subcorpora	735	10194	13.87	9.74	0.31
JEUXVIDEOS.COM	199	3058	15.37	14.44	0.81
TWITTER	216	2465	11.41	7.81	1.24
high noisiness subcorpora	93	1126	12.11	8.51	1.46
other subcorpora	123	1339	10.89	7.20	1.08
FACEBOOK	452	4200	9.29	8.17	1.67
high noisiness subcorpora	120	1012	8.43	7.12	2.44
other subcorpora	332	3188	9.60	8.49	1.30

Table 1: Corpus properties

French linguistic modeling and statistical parsing, were obtained by looking for slang words and urban idiomatic constructions. Table 1 presents some properties of our corpora.

In order to quantitatively assess the level of noisiness in our corpora we defined an ad-hoc *noisiness* metric. It is defined as a variant of the Kullback–Leibler divergence between the distribution of trigrams of characters in a given corpus and that in a reference corpus (in our case, the French Treebank (Abeillé et al. [1]), hereafter FTB). The figures given in Table 1 are consistent with our classification in two noisiness levels. We used this metric to decide for each sub-corpus whether to apply a standard pre-annotation or a dedicated noise-tolerant architecture instead (cf. Section 5).

We refer the reader to (Seddah et al. [13]) for more detail on our various subcorpora. We provide here two examples, (1) from the lightly edited TWITTER subcorpus, and (2), from the our high-noisiness FACEBOOK subcorpus.

(1) Je soupçonnes que "l'enfarineuse" était en faite une cocaineuse vu la pêche de #Hollande ce soir à #Rouen.
Je soupçonne que l'enfarineuse était en fait une cocaïneuse vu la pêche de #Hollande ce soir à #Rouen.
I suspect that the "flouring-lady" was actually a cocaïn-lady given the energy of #Hollande this night at #Rouen.

(2) L'Ange Michael vraiment super conten pour toi mé tora plus grace a moi tkt love you!
L'Ange Michael: (Je suis) Vraiment super content pour mais tu auras plus grace à moi. Ne t'inquiètes pas. Je t'aime !
The Angel Mickael: (I am) Really very happy for him but you'll get more because of me. Don't worry. I love you!

3 Linguistics of user generated content

User-generated texts do not correspond to a single homogenous domain, although some specificities of user-generated content are found across various types of web data. Moreover, in some cases, and most notably TWITTER, such data include both linguistic content and media-specific meta-language. This meta-language (such as

Phenomenon	Attested example	Std. counterpart	Gloss
	Ergographic phenomena		
Diacritic removal	*demain c'est l'***ete**	*demain c'est l'été*	'tomorrow is summer'
Phonetization	*je suis* **oqp**	*je suis occupé*	'I'm busy'
Simplification	*je* **sé**	*je sais*	'I know'
Spelling errors	*tous mes* **examen**	*tous mes examens*	'All my examinations
	son *normaux*	*sont normaux*	are normal'
	Transverse phenomena		
Contraction	**nimp**	*n'importe quoi*	'rubbish'
	qil	*qu'il*	'that he'
Typographic diaeresis	**c a dire**	*c'est-à-dire*	'namely'
	c t	*c'était*	'it was'
	Marks of expressiveness		
Punctuation transgression	*Joli !!!!!!*	*Joli !*	'nice!'
Graphemic stretching	*superrrrrrrrr*	*super*	'great'
Emoticons/smileys	*:-)*, <3	–	–

Table 2: A few idiosyncrasies found within French user-generated content

TWITTER's "RT" ("Retweet"), at-mentions and hashtags) is to be extracted before parsing *per se* or other types of linguistic processing. In this work, we focus on the linguistic content. Therefore, we deal with meta-language tokens only when they are embedded within or adjacent to purely linguistic content (e.g., the tweet itself, provided it consists of one or several sentences).

Prevalent idiosyncrasies in user generated content can be characterized on two axes: one which can be roughly describe as "the encoding simplification axis" which covers ergographic[1] and transverse phenomena and the other "sentiment expression axis" which covers phenomena, or marks of expressiveness, that emulate the same goal as sentiment expressed through prosody and gesture in direct interaction. Table 2 gathers the most striking of these phenomena.

These artifacts lead to a high unknown word level. More importantly, the new morphology brought by the those phenomenon complicates any suffix-based unknown word analysis. Nevertheless, our general annotation strategy consists in staying as consistent as possible with the FTB guidelines (Abeillé et al. [1]).

4 Annotation scheme

We followed the FTB annotation guidelines (Abeillé et al. [1]). More precisely, we based our annotation scheme on its FTB-UC variant (Candito and Crabbé [3]) which was optimized for parsing purposes. It mainly departs from the original FTB on the tagset granularity and on the modeling of multiword units. We added specific guidelines to handle idiosyncrasies user-generated content corpora.

We also added two new POS tags, namely *HT* for TWITTER hashtags and *META* for meta-textual tokens, such as TWITTER "RT". TWITTER at-mentions as well as URLs and e-mail addresses have been tagged *NPP*. The rationale for

[1] Phenomenon aiming at reducing the writing effort.

this is to remain consistent with our tagging and parsing models trained on the FTB, which do not contain such tokens. This constitutes the main difference with other works on user-generated data (Gimpel et al. [11]). One other major extension at the POS level concerns contraction and typographic diaeresis phenomena (see Section 3). Contracted tokens are associated with a combined POS tag which lists the sequence of each underlying words' tag. Let us consider for example, the non-standard contraction *jai*, which stands for *j' ai*, which would have been tagged *CLS* and *V* (subject clitic and finite verb). The non-standard contracted token *jai* is then tagged *CLS+V*. In this case, the contraction involves a verb and one of its argument. In such situations, function labels are associated directly with the contracted token. For cases of typographic diaeresis, the category of its standard counterpart is given to the last token, all others receive the special tag *Y*. For example, *c a dire* stands for the conjunction *c'est-à-dire*, which would have been tagged *CC*. We thus tag the first two tokens as *Y* and *dire* as *CC*. This is consistent with how such cases are handled in the English Web Treebank (Bies et al. [2]).

At the syntactic level, the main addition to the FTB-UC tagset is a new FRAG label, for phrases that cannot be syntactically attached to the main clause of a syntactic unit (e.g., salutations, emoticons...). It also covers usernames, at-mentions, and URL appended to a sentence.

These extensions are largely compatible with the English Web Bank. However, our treebank differs from the former in several aspects. First, French has a richer morphology than English, entailing a tedious disambiguation process when facing *noisy* data. Although the first version of our treebank is smaller than the English Web Treebank, it includes richer annotations (compound POS, corrected token form of contractions) and includes subcorpora exhibiting a very high level of noise.

5 Annotation Methodology

We built our manually validated treebank following a well established methodology: we first defined a sequence of annotation layers, namely (i) sentence splitting, tokenization and POS tagging, (ii) syntagmatic parsing, (iii) functional annotation. Each layer is annotated by an automatic preprocessing that relies on previously annotated layers, followed by validation and correction by human annotators. At each step, annotators were able to modify choices made at previous stages.

We used two different strategies for tokenization and POS pre-annotation of our sub-corpora, depending on their noisiness score. For less *noisy* corpora (noisiness score below 1), we used a slightly extended version of the tokenization tools from the FTB-based parsing architecture Bonsai (Candito et al. [4]), in order to match as much as possible the FTB's tokenization scheme. Next, we used the POS-tagger MORFETTE (Chrupała et al. [5]). For corpora with a high noisiness score, we used a specifically developped pre-annotation process. This is because in such corpora, spelling errors are even more frequent, but also because the original tokens rarely match sound linguistic units. The idea underlying this pre-processing is to wrap

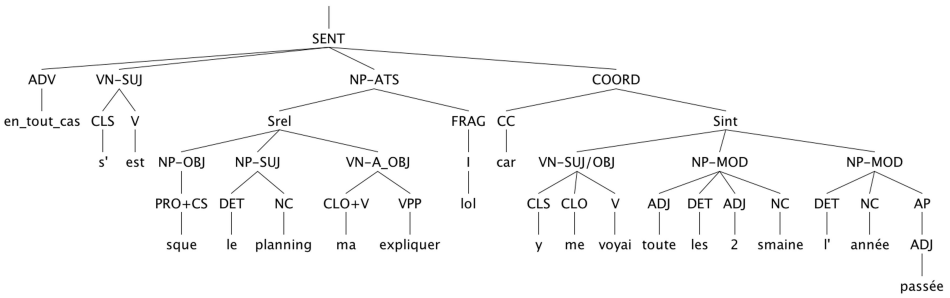

Figure 1: French Social Media Bank's sample of the *noisy*DOCTISSIMO subcorpus. English gloss: 'Anyway that's what the social centre explained to me lol he was actually seeing me every two weeks last year.'

the POS tagger (in this case, MElt, (Denis and Sagot [7])) within a temporary text normalization tool, so that the tagger is provided with data as close as possible to its training corpus, the FTB.

Parse pre-annotation was achieved using a state-of-the-art statistical parser trained on the FTB-UC, provided with manually corrected POS tags. We used the Berkeley parser (Petrov et al. [12]) adapted to French (Crabbé and Candito [6]). Note that when the validated POS tags were discarded by the parser, in case of too many unknown word-POS pairs, those were reinserted. Functional annotation was carried out as a post-parsing stage using the associated labeler (Candito et al. [4]) and then manually validated. An example of the resulting annotation is shown Figure 1.

6 Conclusion

The French Social Media Bank shares with the English Web Treebank (Bies et al. [2]) a common will to extend the treebank domain towards user generated content. Although of a smaller scale, it constitutes one of the very first resources for validating social media parsing and POS tagging, together with DCU's Twitter & BBC football forum treebanks (Foster et al. [9, 10]) and the Twitter POS data set from Gimpel et al. [11]. Moreover, it is the first set of syntactically annotated FACEBOOK data and the first treebank of its kind for French.

We performed a first round of evaluation showing that simple techniques could be used to improve POS tagging performance. Indeed, raw accuracy results of the MElt POS-tagger, which gives state-of-the-art results on edited texts, range from 56 % (DOCTISSIMO-noisy) to 87 % (TWITTER), whereas the use of the dedicated wrapper mentioned in Section 5 leads to figures between 80 % and 89 %. We have also achieved baseline statistical parsing results, with results far behind those on newspaper in-domain texts (Evalb's f-measures ranging from 39 % to 70 %, to be compared with 86–89 % regularly achieved on the FTB test set). These preliminary results prove the difficulty of processing such data and therefore the importance of building a data set such as the French Social Media Bank.

Acknowledgments This work was partly funded by the French ANR project EDyLex (ANR-09-CORD-008).

References

[1] Abeillé, A., Clément, L., and Toussenel, F. (2003). *Building a Treebank for French*. Kluwer, Dordrecht.

[2] Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English web treebank. Technical report, Linguistic Data Consortium,, Philadelphia, PA, USA.

[3] Candito, M. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of IWPT'09*, Paris, France.

[4] Candito, M., Nivre, J., Denis, P., and Henestroza, E. (2010). Benchmarking of statistical dependency parsers for french. In *Proc. of CoLing'10*, Beijing, China.

[5] Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco.

[6] Crabbé, B. and Candito, M. (2008). Expériences d'analyse syntaxique statistique du français. In *Proc. of TALN'08*, pages 45–54, Senlis, France.

[7] Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proc. of PACLIC*, Hong Kong, China.

[8] Foster, J. (2010). "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Proc. of HLT/NAACL'10*, Los Angeles, USA.

[9] Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and van Genabith, J. (2011a). #hardtoparse: Pos tagging and parsing the twitterverse. In *Proc. of the AAAI 2011 Workshop On Analyzing Microtext*.

[10] Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Nivre, J., Hogan, D., and van Genabith, J. (2011b). From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *proc of IJCNLP*, Chiang Mai, Thailand.

[11] 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 *Proc. of ACL'11*, Portland, USA.

[12] Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proc. of ACL'06*, Sydney, Australia.

[13] Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The french social media bank: a treebank of noisy user generated content. In *Proceedings of CoLing'12*, Mumbai, India.