Normalizing tweets with neural language models

Grzegorz Chrupała
g.chrupala@uvt.nl

ABSTRACT

Due to size limitations and genre conventions tweets often contain a large proportion of abbreviations, alternative spellings, novel words and other non-canonical language. These features are problematic for standard language analysis tools, so it can be desirable to normalize tweets, i.e. convert them to canonical form.

- Align original and normalized strings at character level
- Find the shortest edit script which transforms original into normalized
- Treat the edit operation at each position in original string as a label
- Use a Conditional Random Field as a model to learn such labels.
- Use character ngrams as features
- Use learned text embeddings as additional features

The text embeddings are generated using an Simple Recurrent Network (aka Elman Net) as a language model. The embedding at a certain position in a string is the activation of the hidden layer as the network is predicting the character at this position.

The neural language model was train on a raw sample of tweets (414 million bytes), without language-based or any other type of filtering.

We find that enriching the feature set with learned text embeddings substantially lowers word error rates on tweet normalization on two datasets in two languages.

NORMALIZATION EXAMPLES

I will c wat i can do
i will see what i can do

imma jus start puttn it out there
imma just start putting it out there

Buuenoo poos mee voii muu contentaa
Bueno pues me voy muy contenta

Hechoo de menos a mi mami :'(
Echo de menos a mi mamá :'(

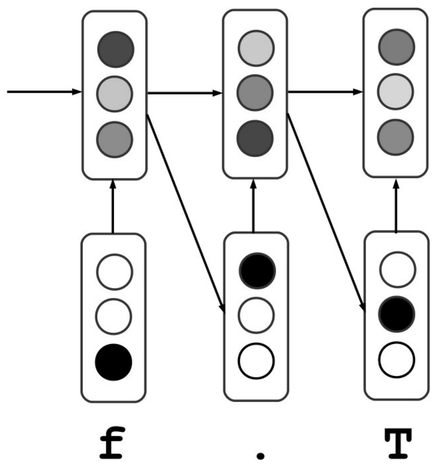
ALIGNMENT

Original	Deletions	Insertions	Label
w			NULL
i			NULL
l			NULL
l			NULL
			NULL
c	c		DELETE
		see	INSERT(see)
w			NULL
a		h	INSERT(h)
t			NULL

N-GRAM FEATURES

Position		Unigrams	Bigrams	Trigrams
-1	w			
0	**a**	**w a t**	**wa at**	**wat**
+1	t			

ELMAN NET EMBEDDINGS



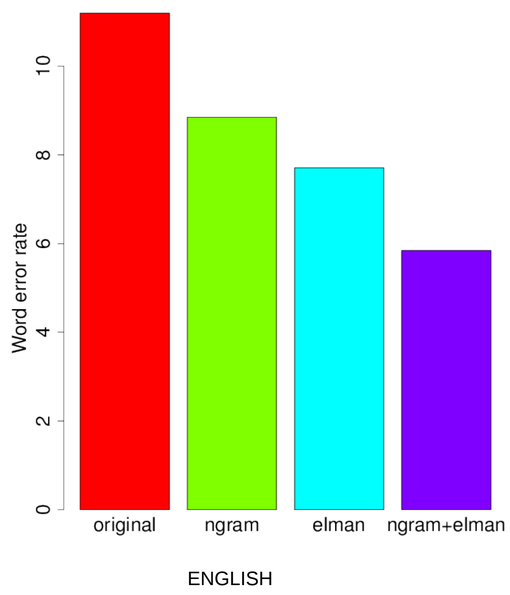
Example nearest neighbors in embedding space

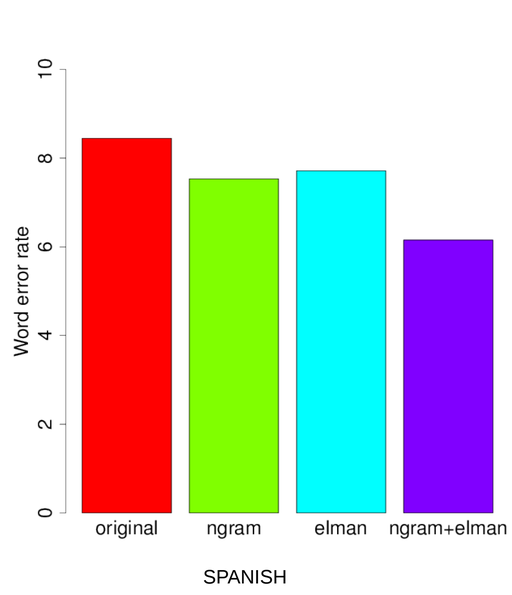
should h	should d	will s	will m	should a
@justth	@neenu	@raven_	@lanae	@despic
maybe u	maybe y	cause i	wen i	when i

RANDOMLY GENERATED TWEETS

@YuszLAL100A 暇すぎるｗｗｗｗｗとか麺役者についてる…(ヽ >
晒せ 信じに行けていいんだな…。 RT @yaepdrrafa:
@fsch_chany siaaa,, dobek taha subus sama kiri kabur wanak... hahah
なかなかない。
やばい
But I'm the good first-Good Chulc

RESULTS





REFERENCES

Grzegorz Chrupała. 2013. Text segmentation with character-level text embeddings. ICML Workshop on Deep Learning for Audio, Speech and Language Processing.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, Johan Bos. 2013. Elephant: Sequence Labeling for Word and Sentence Segmentation. EMNLP.

Bo Han and Timothy Baldwin. 2011. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. ACL.

Iñaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluís Padró, Iñaki San Vicente, Jordi Turmo, Arkaitz Zubiaga. 2013. Introducción a la Tarea Compartida Tweet-Norm 2013: Normalización Léxica de Tuits en Español. SEPLN.