Hashtag Segmentation of Conversational Tweets in Turkish, Modelling Segmentation Approach Midterm Report

Utku Saridede, Sevket Topuz

Informational Retrival and Natural Language Processing, Department of Computer Engineering, Bogazici University, Istanbul Bebek 34342, TR

Email: utku.saridede@boun.edu.tr, sevket.topuz@boun.edu.tr

Twitter is the latest social networking tool which affects everything related to the person. Twitter allows its users to write at most 140 character long update, it is known off as "tweet". In this research, analyzing segmentation of tweets is our main objective. There are several researches in English, but not that much in Turkish. Studying with the Turkish corpus is somehow hard to handle, because Turkish resources are insufficient. The usage of the hashtags differ in country to country, that is, some countries do not know off proper usage of hashtags. Having more than one word in the hashtag or lapsus calami are misusage, and prevent researhes to work properly. The other part of the project is analyzing hashtags in the way of linguistics. Analyzed corpus will give more information about related countries. In the other words, short-term hashtag analyses keep informed about spesific situations which influence the society.

Keywords: Twitter; Tweets; Tweet; Hashtag; Segmentation; Turkish; Conversational Tweets; Hashtag Segmentation

Received 07 October 2016; revised 08 November 2016

1. INTRODUCTION

A hashtag is defined by any string prefixed with a "#", for instance, #freedomtomark, #shesuggest. The string can be a single word, an acronym, or multiple words joined together, and usually identifies the subject topic of the tweet (e.g., #ENG493) or expresses a comment about it (e.g., #kappamevku).

The main objective is to implement machine learning based hashtag segmentation application. Our first work was using twitter developer tools to extract tweets from their database. In the case of hashtag extraction, there are several issues. For instance, we have recognized that Turkish people do not understand the usage of hashtag approach. Therefore, the formation of corpus becomes difficult. Using large amount of raw data that

is recieved from social media is the way of creating corpus. In the field of natural language processing, the main requirement is datasets. It clarifies why there is not enough research in Turkish. Improving datasets might help researchers to test their idea and modellings.

The starting point of project is mainly based upon gathering data to avoid backing to drawing point. That is, being blind to quantity of data causes to quit idea. Existing methods of word segmentation unsuperwised language models. Researches claim that using multiple corpora, the joint probability model from multiple corpora performs significantly better than the individual corpora. Weighted joint probability model, with weights is specific to each corpus. Decent approach is to train the weights in a supervised manner using max-margin methods, that is, a machine learning method to make a decision about the boundries of words. The supervised probability models improve segmentation accuracy over joint probability models. Researchers observe that length of segments is an important parameter for word segmentation, and incorporating length-specific weights into our model supports the current model. However the length specific models further improve segmentation accuracy over supervised probability models.

All mentioned models try to solve the problems with dynamic programming algorithms. The supervised length specific models have significantly more advencement over unsupervised single corpus and joint probability models. Segmentation of hashtags result in significant improvement in recall on searches for twitter trends.

2. STATE OF ART

Word segmentation are of great interest to Natural Language Processing researchers. A good number of methods have been proposed in the literature, with quite good performances reported. There are several articles about word segmentation. There are two major categories, as well as, boundary prediction and word recognition. One of the most frequently used method for that is maximum matching, by Wong and Chan 1996. Many approaches help us to handle unknown words and to get best proper answer in the case of ambiguity.

Boundary prediction methods usually utilize local statistics to decide whether there is a word boundary between two language units given the local context. The representative examples involve Ando-Lee Criterion (Ando and Lee 2000), Mutual Information (Sun, Shen, and Tsou 1998) and Branching Entropy (Jin and Tanaka-Ishii 2006). Recently Fleck (2008) proposed a algorithm called WordEnds. It trained a boundary classifier with the dixit boundary cues and then used it to mark word boundaries. Zhikov, Takamura, and Okumura (2010) proposed an efficient algorithm combining the strength of Minimum Description Length approach and local statistics Branching Entropy. High performance in terms of both accuracy and speed was reported.

Word boundry detectiton and word segmentation is very important for Chinese words research. A good survey about Chinese word segmentation can be found out in Wu and Tseng's paper. A Chinese senteces do not include delimiters to seperate words. It includes composed of a string of characters. Neural networks and lazy learning approaches are methods that are used in word segmentation.

Another approach is word segmentation of URL links. To think about the word segmentation and recognition problem for URL links, reasearchers adopt some basic principles from rulebased approach. The dictionary should have sufficient amount of word entries. However, the occurence of compound words makes it very difficult to match every component string exactly with the dictionary entries. Evaluation of the hashtag segmentation has started to be improved with search engine improvements of Twitter.

3. METHODS

We have tried to discourse manly with 3 methods. Hashtag segmentation can be generally defined as word boundry detection. Because of this, we start with detection of the word boundry. There are two featurebased learning methods, Conditional Random Fields (CRFs)(Laerty et al.,2001) and Maximum Entropy (MaxEnt). CRFs can represent the uncommon parts of the information as elements furthermore, are great at displaying grouping marking problems. MaxEnt is extremely compelling at learning with a high assortment of components, without agonizing over the multifaceted nature of the model. Hidden Markov Model is a simplistic approach for word segmentation. It helps us to built character trigrams. It tries to catch boundary characters that are current and previous ones. Peter Norvig's implementation can be used for word bigrams.

Manual annotation is time consuming task and it limits the amount of trainig data that can be created. We try to achieve utilizing data to create training sets for hashtag segmentation. Synthetic hashtags by concatenating the words in tweets can also be used for training data because word boundries are known. To use concatenating the words in tweets as training dataset, we need to filter nonword tokens. If tweets include nonword token in the beginnig or end of the text, it can be removed and other words can be used as trainig data. On the other side, if a nonword token appear in the middle of the text, the tweet is dicarded because nonword token ib the middle of the tweet may distort the word order. The word order is important point of trainig data.

We can use each character of training data to represent one function of learning system. Some features should be determined and each character should be examined according to these features to create machine learning system.

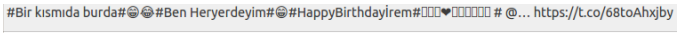


FIGURE 1. Tweet example.

	ID	BODY
10	663003460036620288	RT @SAkyol75: Kaç kardeşsiniz dediklerinde, Bir buçuk milyar diyorum. Anlatabiliyormuyum ? #Sezai Karakoç # @sed
11	663003815977877504	Elma sekeri yaptımmm isteyen var mii❤□□□□□#birkapkek # elmasekeri... https://t.co/UfkJCjs3G1
12	663004961358376960	# Muslera Sabri Chedjou Balta Olcan Podolski Selçuk Sneijder Yasin Burak Umut
13	663005182662438912	RT @AsiklariFener: A Milli Takım Aday kadrosuna Emre Çolak ve Yasin Öztekin alınırken Volkan ŞEN alınmadı! # Fener
14	663005506039095296	#Hayata gülümsemek #
15	663005521344004096	#incil #tevrat Hristiyanlık'a götüren İslamiyet https://t.co/tUH2C6SO19 # #zebur #kuran https://t.co/3zBkCGmRvr
16	663005528130383872	Hristiyanlık'a götüren İslamiyet https://t.co/OLiUdGQPo1 # #allah #islam https://t.co/XgkpJ0nHln
17	663005544567844864	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/z33WCH1F1e # #hristiyan https://t.co/NaDyt4RsDo
18	663005557033295872	#isa #mesih Hristiyanlık'a götüren İslamiyet https://t.co/Blev4zUwqi # #muhammed https://t.co/9NFExsQhJg
19	663005565736456193	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/TAtaRaqrva # #hristiyan https://t.co/mC0cszKzyX
20	663005579456065536	Hristiyanlık'a götüren İslamiyet https://t.co/PnJIEZnNKp # #türk #tanrı #kilise https://t.co/Cvp3p9M5Vh
21	663005586921926656	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/TAtaRaqrva # #hristiyan https://t.co/mC0cszKzyX https://t.c
22	663005951004385284	# bebrillant@htcchampions https://t.co/3uu9DVlhDW

FIGURE 2. Text database table representation.

4. RESULTS

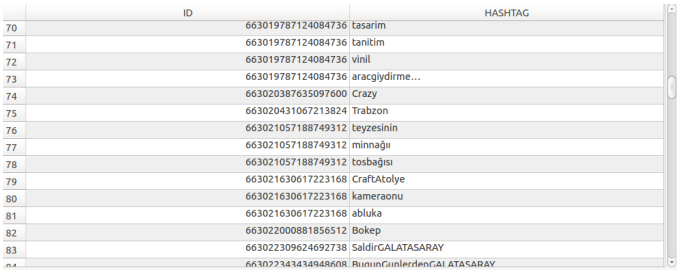
The training data will be the part of the current data. We tokenize tweets, normalize them and get hashtags from them.

Implementing a method to get hashtags and insert them database is accomplished. There are two kind of table as well as "TEXTS" which contains unique tweet IDs and tweet text.

The other one is "HASHTAGS" which contains tweet IDs and hashtags. Segmentation algorithm is almost done.

5. CONCLUSION

We proposed a simple and effective unsupervised word segmentation approach. The criterion incorporates boundary information to model words.



	ID	HASHTAG
70	663019787124084736	tasarim
71	663019787124084736	tanitim
72	663019787124084736	vinil
73	663019787124084736	aracgiydirme...
74	663020387635097600	Crazy
75	663020431067213824	Trabzon
76	663021057188749312	teyzesinin
77	663021057188749312	minnağı
78	663021057188749312	tosbağısı
79	663021630617223168	CraftAtolye
80	663021630617223168	kameraonu
81	663021630617223168	abluka
82	663022000881856512	Bokep
83	663022309624692738	SaldirGALATASARAY

FIGURE 3. Hashtag database table representation.

6. FUTURE WORK

We decided to improve our segmentation algorithm. Features of the words will be dicussed. Training and real datasets will be prepared. Every word in the training data will be used to improve machine learning approach of our project. Later future works will be consulted with lecturer.

7. REFERENCES

- Bansal, P., Bansal, R. and Varma V. 2015. Towards Deep Semantic Analysis Of Hashtags. To Appear in 37th European Conference on Information Retrieval

- Berardi, G. and Esuli, A. and Marcheggiani, D. and Sebastian, F. 2011. ISTI@TREC Microblog track 2011: exploring the use of hashtag segmentation and text quality ranking.. In The Twentieth Text RE- trieval Conference Proceedings

- Chen, S. and Xu Y. and Chang, H. 2012. A Simple and Effective Unsupervised Word Segmentation Approach. Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

- Xue, N. 2003. Chinese word segmentation as character tagging.. International Journal of Computational Linguistics and Chinese Language Processing vol- ume 8(1)

- Wong, P and Chan, C. 1996. Chinese word segmenta- tion based on maximum matching and word binding force. In the proceedings of the 16th conference on Computational linguistics V(1) 200-203