

---

# Hashtag Segmentation of Conversational Tweets in Turkish, Modelling Segmentation Approach Final Report

UTKU SARIDEDE, SEVKET TOPUZ

ADVISOR: ASS. PROF. ARZUCAN OZGUR

CO-ADVISOR: ARDA ELEBI

*Degree of Bachelor of Science*

*Natural Language Processing and Informational Retrival, Department of Computer*

*Engineering, Bogazici University, Istanbul Bebek 34342, TR*

*Email: utku.saridede@boun.edu.tr, sevkettopuz@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 hashtags from 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 study, because Turkish resources have grammer problems due to the English effect. The usage of the hashtags differ in country to country. Having more than one word in the hashtag or lapsus calami prevent researchers to work properly. It is the first project in Turkey to segment tweet’s hashtags in Turkish. Therefore, the results of our project is milestone in that manner. It will assist oncoming projects in the case of corpus and method needs. 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. After all said and done, we will recognize the strength of computer science on natural languages.

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

*Received 07 October 2015; revised 12 January 2016*

**CONTENTS**

<b>1</b>	<b>Introduction and Motivation</b>	<b>3</b>
1.1	Introduction . . . . .	3
1.2	Motivation . . . . .	4
<b>2</b>	<b>State of Art</b>	<b>4</b>
2.1	Related Papers and Projects . . . . .	4
2.2	Improvement of Our Project . . . . .	7
<b>3</b>	<b>Methods</b>	<b>7</b>
3.1	Current Methods . . . . .	7
3.2	Our Method . . . . .	8
<b>4</b>	<b>Results</b>	<b>8</b>
<b>5</b>	<b>Conclusion</b>	<b>8</b>
<b>6</b>	<b>Future Work</b>	<b>9</b>
<b>7</b>	<b>References</b>	<b>9</b>
<b>8</b>	<b>Appendix</b>	<b>9</b>

## 1. INTRODUCTION AND MOTIVATION

### 1.1. Introduction

#### 1.1.1. The Story of Hashtag

A hashtag is a type of label or metadata letters used especially on social network and microblogging services which makes it easier for users to find messages with a specific theme or content.

The story is began with Twitter but has extended to other social media platforms as well as “facebook”. In 2007, developer Chris Messina proposed, in a tweet, that Twitter begin grouping topics using the hash symbol. Twitter initially rejected the idea. But in October 2007, citizen journalists began using the hashtag “#SanDiegoFire”, at Messinas suggestion, to tweet updates on a series of forest fires in San Diego.

#### 1.1.2. Decision of Hashtags

Which characters can be defined as #hashtag is the important part of our project.

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).

Spaces are an absolute segmentation rule. Even if hashtag contains multiple words, they should be together. Using capital letters in between words have no meaning. (#CahitArf). Uppercase letters will not alter search results, so searching for #CahitArf will yield the same results as #cahitarf.

Numbers are supported in Twiter, so as #23NisanBayrami. However; punctuation marks, commas, periods, exclamation points, question marks and apostrophes are forbidden characters. In addition to them; asterisks, ampersands or any other special characters are also restricted ones. There is no preset list of hashtags. Creating a brand new hashtag is simple by putting the hash before a series of words, and if it hasn’t been used before, a new hashtag is invented.

#### 1.1.3. Origin of Implementation

The main objective is to implement machine learning based hashtag segmentation application.

The first work was using twitter developer tools to extract tweets from Tweeter’s database. In the case of hashtag extraction, there are several issues. 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 essential requirements are datasets. Having realiable training and test data helps to improve current algorithm. Because languages are flexible and few training and test data cause to reproduce wrong idea. It somehow clarifies why there is not enough research in Turkish. That is because, improving datasets might help researchers to test their idea and models in the future.

The starting point of project is primarily based upon gathering sufficient data to avoid backing to drawing point. That is, being blind to quantity of data causes to quit idea. Hence, the researcher should collect large amount of data, but also with well-selected contents. When the research topic comes to natural language processing, size and quantity of data is important. Extending corpus enhances current models to achieve better results.

#### 1.1.4. Word Segmentation

Word segmentation means dividing a text into meaningful words. Human-beings can divide text into words with their mental process, but computer not. Word segmentation became more difficult, meanwhile not using a separator or using more than one form for separation. Hence, natural language processing begins after that field.

There are several methods about word segmentation. Methods will be discussed in the case of convenience with Turkish.

### 1.2. Motivation

The common problem about languages which have Latin alphabet system is deciding the word boundary. There are three types of word boundary type.

First one is using space between words. We can not use space segmentation in our project, because hashtags do not contain spaces. Second type is using uppercase letters at the beginning of words or using underscore between the words. It is the main separation rule of our word segmentation. If hashtags have more than one word in it, we can check the uppercase letters or underscores to decide word boundaries. However; when it comes to real world, the usage of letters in hashtags differs from the second type. The third one is using no uppercase or using nonsense uppercases in hashtags. Because of the natural languages' aspects, datasets contain the hashtags which are the third type of boundary type.

For natural language processing, we have to determine the word's boundaries first. The method that we used tries to work on collected data to create a model that demonstrates the Turkish words' structures.

## 2. STATE OF ART

### 2.1. Related Papers and Projects

There are no paper or research about hashtag segmentation in Turkish. So that, the discussion will be based upon other versions of hashtag segmentation and as well as word segmentation.

The problem of word segmentation has been studied in various contexts. 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. Venkataraman<sup>1</sup> uses unsupervised word segmentation techniques for finding words in automatically transcribed speech. Recently, Macherey et al.<sup>2</sup> use unsupervised techniques for word segmentation, highlighting the application of this technique to multiple European languages. The work of Wang et al.<sup>3</sup> use unsupervised techniques with multiple corpora for word segmentation.

#### 2.1.1. Exploring the Use of Hashtag Segmentation and Text Quality Ranking

A common practice in tweets is to identify their subject topic by means of a hashtag. In this project, they have given information that a hashtag cannot contain white spaces. They also noticed that people usually concatenate

---

<sup>1</sup>A. Venkataraman. A statistical model for word discovery in transcribed speech. Computational Linguistics, 2001.

<sup>2</sup>K. Macherey, A. M. Dai, D. Talbot, A. C. Popat, and F. Och. Language-independent compound splitting with morphological operations. In ACL HLT, 2011.

<sup>3</sup>K. Wang, C. Thrasher, and B.-J. P. Hsu. Web scale nlp: a case study on url word breaking. In Proceedings of the 20th international conference on World wide web, New York, NY, USA, 2011. ACM.

more words together, to form a short phrase. They also mentioned that the distinct words composing a hashtag is not a simple task to be automatized.

It is obvious that our project and their project are based on the same aspects like; “Some users identify the distinct words using a CamelCase style, i.e., capitalizing the first letter of each word, other leave the words all in lowercase or uppercase. Other use underscores \_ to separate words, but it is not a common case because it wastes characters. Usually words are just juxtaposed without any evident or coherent use of separation signs and actually there are no common rules one can rely on to segment hashtags.”<sup>4</sup> They demote their problem as word segmentation problem. The next figure represents the working principles of their system. That system uses the same learning data path as we have.

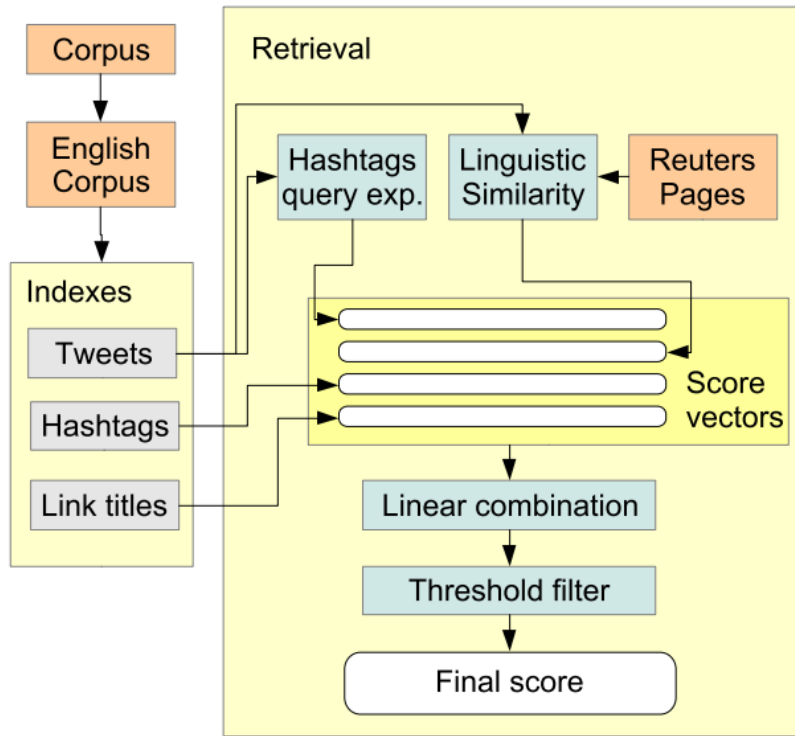


FIGURE 1. Modules of System

They used the words distribution model and a hashtag, the hashtag segmentation module converts the hashtag to a vector of words composing them. Finally, the last figure related to that paper is about how big dataset’s size is.

### 2.1.2. Segmenting Web-Domains and Hashtags using Length Specific Models

In this project, they study two applications in the internet domain. First application is the web domain segmentation which is crucial for monetization of broken URLs. Secondly, they propose and study a novel application of twitter hashtag segmentation for increasing recall on twitter searches. They remark that existing methods for word segmentation use unsupervised language models.

They have realized that when using multiple corpora, the joint probability model from multiple corpora performs significantly better than the individual corpora during the project. Motivated by this, they propose weighted

<sup>4</sup>ISTI@TREC Microblog track 2011: exploring the use of hashtag segmentation and text quality ranking, Giacomo Berardi, et al.

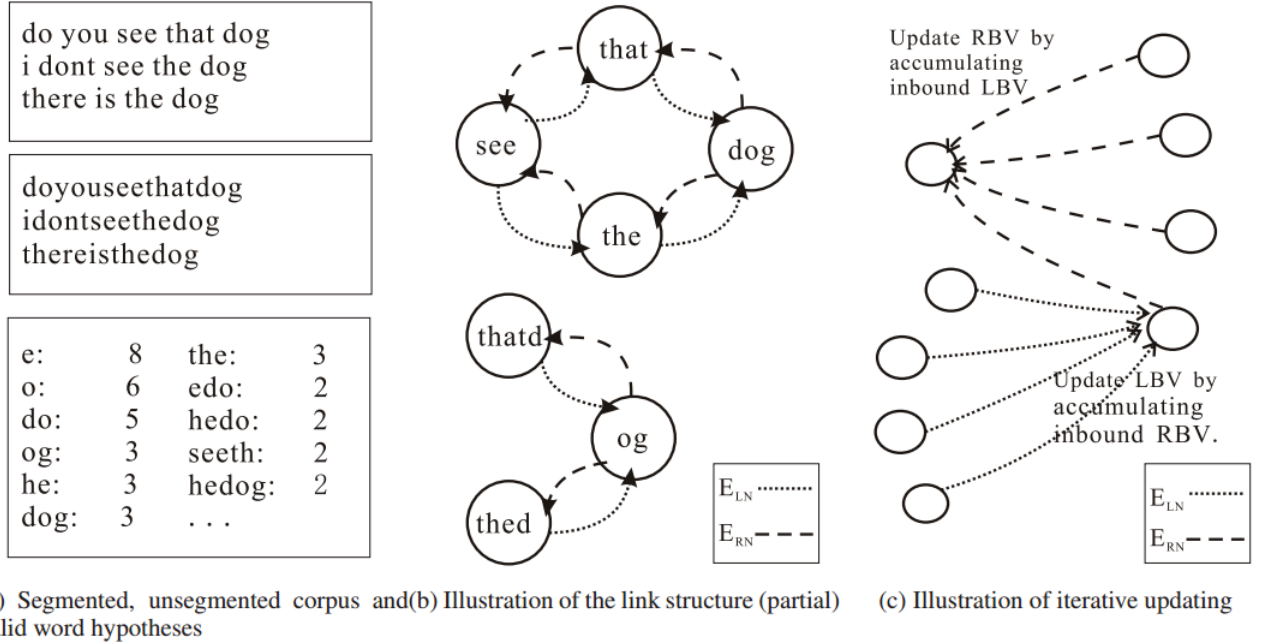
	total	effective	retweets	null	hashtags	users
Entire corpus	16.141.812	13.812.346	1.104.780	1.224.686	655.850	5.356.842
English set	4.510.329	4.068.158	442.171		288.753	2.021.759
Hashtags subset	791.464	640.870	150.594		288.753	514.401
Link subset	676.957	674.471	2.486		14.399	400.631

**FIGURE 2.** Some statistics from the corpus and the subsets that is selected for indexing, the total number of tweets is divided in effective tweets, retweets and null tweets. The hashtags column indicates the number of unique hashtags.

joint probability model, with weights specific to each corpus. Finally, they observed that length of segments is an important parameter for word segmentation. The length specific models further improve segmentation accuracy over supervised probability models.

### 2.1.3. A Simple and Effective Unsupervised Word Segmentation Approach

In this paper, they propose a new unsupervised algorithm to achieve word segmentation. The main idea of their approach is a novel word induction criterion which is similar with second paper. As they explain; “We devise a method to derive exterior word boundary information from the link structures of adjacent word hypotheses and incorporate interior word boundary information to complete the model.”<sup>5</sup> They have also worked with Chinese datasets to claim their approach. If it is working with even difficult language systems, it will be achieved with other systems. They also reported that their approach is simpler and more efficient than the Bayesian methods and more suitable for real-world applications.



**FIGURE 3.** Illustrations of constructing the link structures of word hypotheses and calculating the exterior boundary values.

<sup>5</sup>A Simple and Effective Unsupervised Word Segmentation Approach, Songjian Chen, Yabo Xu, Huiyou Chang

#### 2.1.4. URL Segmentation

Another approach is word segmentation of URL links. To think about the word segmentation and recognition problem for URL links, researchers adopt some basic principles from rulebased approach.

The dictionary should have sufficient amount of word entries. However, the occurrence 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.

Finite state transducer is an approach for title token base URL segmentation. It splits and segments according to previous-seen web page title simultaneously. For example, URL link includes "cs" that might correspond to "computer science" in many training page's title. If "cs" is encountered in the testing corpus, it is automatically expanded to "computer science". The transducer has several rules that give score. This score can be obtained by certain moves which match or skip letters in the tokens of title with corresponding letters. Expansion can be valid if it covers all letters in the segment.

## 2.2. Improvement of Our Project

In our project, there are several improvements according to local and global researches. Firstly, as we mentioned before, it is the first project that introduces a platform for segmentation of Twitter hashtags in Turkish. We also gathered datasets from Twitter and parse them into "tweet body" and "hashtags". We have used maximum entropy model to separate words. We have also implemented an application feature to embed project into any platform. If we get hashtags, we can return the segmented version of them

## 3. METHODS

### 3.1. Current Methods

We will discuss two types of models and both models focus on words, not boundaries. And they also use little or no domain-specific information.

#### 3.1.1. Parser Model

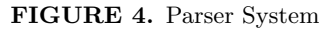
No special mechanism is needed for word segmentation; it results from interaction of perception and internal representation. Humans are not tracking boundary statistics; segmentation results from general properties of attention, perception, and memory.

Initially, input is perceived and chunked randomly into units;

- Units are encoded in memory.
- Memory decays rapidly.
- Uncommon units disappear, common units are reinforced.
- Units in memory influence perception and encoding of new input (input is segmented into existing units).

The properties of parser model;

- No explicit tracking of statistics is needed.
- Works on experimental stimuli but might need modifications for realistic language.
- Probably would work in many domains.



## SARIDEDE, TOPUZ



ID	BODY
663003460036620288	RT @SAkyol75: Kaç kardeşinizin dediklerinde, Bir buçuk milyar diyorum. Anlatabiliyormuyum ? #Sezai Karakoç # @sed
663003815977877504	Elma sekeri yaptımın isteyen var mı? #birkapkek # elmasakeri... https://t.co/UfLJCjs3G1
663004961358376960	# Muslera Sabri Çetioğlu Balta Olcan Podolski Selçuk Şenelider Yasın Burak Ümit
663005182662438912	RT @AsiklarFener: A Milli Takım Aday kadrosuna Emre Çolak ve Yasin Öztekin alınırken Volkan ŞEN alınmadı # Fener
663005506039095296	#Hayata gülümsemek #
663005521344004096	#İncil #tevat. Hristiyanlık'a götüren İslamiyet https://t.co/UH2C6SO19 # zebur #kuran https://t.co/3zBKCGmRvr
663005528130383872	Hristiyanlık'a götüren İslamiyet https://t.co/OLiudGQp01 #allah #islam https://t.co/XgkpJ0nHln
663005544567844864	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/z33WCH1F1e # hristiyan https://t.co/NaDyt4RsDo
663005557033295872	#isa #mesih Hristiyanlık'a götüren İslamiyet https://t.co/Blev4zUwqi # #muhammed https://t.co/9NFExsQhJg
663005565736456193	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/TataRaqrva # #hristiyan https://t.co/mC0cszKzyX
663005579456065536	Hristiyanlık'a götüren İslamiyet https://t.co/PnJIEZnNkp # #türk #tanrı #kilise https://t.co/Cvp3p9MSVh
663005586921926656	#hristiyanlık Hristiyanlık'a götüren İslamiyet https://t.co/TataRaqrva # #hristiyan https://t.co/mC0cszKzyX https://t.co/
663005951004385284	# bebrillant@htccampions https://t.co/3uu9DVLhDW

FIGURE 6. Text database table representation.

ID	HASHTAG
663019787124084736	lasarim
663019787124084736	lanitim
663019787124084736	vinil
663019787124084736	aracglydirme....
663020387635097600	Crazy
663020431067213824	Trabzon
663021057188749312	teyzesinin
663021057188749312	minnağı
663021057188749312	tosbağı
663021630617223168	CraftAtolye
663021630617223168	kameraonu
663021630617223168	abluka
663022000881856512	Bokep
663022309624692738	Saldır GALATASARAY
66302234343494860R	BununGünlerdenCAI ATASARAY

FIGURE 7. Hashtag database table representation.

## 6. FUTURE WORK

We decided to improve our segmentation algorithm. Features of the words will be discussed. Training and test datasets will be prepared. Every word in the training data will be used to improve machine learning power of our project. Later future works will be consulted to Prof. OZGUR.

## 7. REFERENCES

- Berardi, Giacomo; Esuli, Andrea; Marcheggiani, Diego and Sebastiani, Fabrizio. Exploring the use of hashtag segmentation and text quality ranking Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche 56124 Pisa, Italy
- Kan, Min-Yen, Web Page Classification
- Bansal, P., Bansal, R. and Varma V. 2015. Towards Deep Semantic Analysis Of Hashtags.
- 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.
- 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 segmentation based on maximum matching and word binding force.

## 8. APPENDIX