An Analysis of Biomedical Tokenization: Problems and Strategies

Noa P. Cruz Díaz
University of Huelva, Huelva, Spain
noa.cruz@dti.uhu.es

Manuel M. Maña López
University of Huelva, Huelva, Spain
manuel.mana@dti.uhu.es

Abstract

Choosing the right tokenizer is a non-trivial task, especially in the biomedical domain, where it poses additional challenges, which if not resolved means the propagation of errors in successive Natural Language Processing analysis pipeline. This paper aims to identify these problematic cases and analyze the output that, a representative and widely used set of tokenizers, shows on them. This work will aid the decision making process of choosing the right strategy according to the downstream application. In addition, it will help developers to create accurate tokenization tools or improve the existing ones. A total of 14 problematic cases were described, showing biomedical samples for each of them. The outputs of 12 tokenizers were provided and discussed in relation to the level of agreement among tools.

1 Introduction

Tokenization is considered the first step in Natural Language Processing (henceforth, NLP) and it is broadly defined as the segmentation of text into primary building blocks for subsequent analysis (Webster and Kit, 1992).

Tokenization may seem simple if we assume that all it involves is the recognition of a space as a word separator (Baeza-Yates and Ribeiro-Neto, 2011). However, a closer examination will make it clear that a blank space alone is not enough even for general English (Jurafsky and Martin, 2009). Furthermore, choosing the right tokenization strategy is a non-trivial task, especially in the biomedical domain where it poses additional challenges (He and Kayaalp, 2006) which if not resolved means the propagation of errors in successive NLP analysis pipeline. As a consequence, text mining modules, such as Named Entity Recognition, will inevitably suffer in terms of effectiveness (Tomanek et al., 2007).

Tokenization in biomedical literature is particularly difficult due to the fact that general English differ from biomedical text in vocabulary and grammar (Barrett, 2012). In addition, scientific information has a particular structure (Harris, 2002). For example, Campbell and Johnson (2001) carried out three experiments to evaluate the syntactic dissimilarities between medical discharge summaries and everyday English, showing significant differences in syntactic content and complexity.

Another feature of the biomedical literature is related to terminology, which is inconsistently spelt and may vary from typographical errors to lower case and capitalized medication names (Krauthammer and Nenadic, 2004). Furthermore, biomedical texts could be ungrammatical (especially, clinical documents) as well as often include abbreviations and acronyms. Biomedical terms contain digits, capitalized letters within words, Latin and Greek letters, Roman digits, measurement units, list and enumerations, tabular data, hyphens and other special symbols. In addition, another complexity is the ambiguity, i.e., words and abbreviations that have different meanings (homonymy) and concepts described in more than one way (synonymy). For these reasons, the identification of terminology in the biomedical literature is one of the most challenging research topics in the last few years in NLP and biomedical communities and tokenization plays an important role in handling them.

There is no widely accepted tokenization method for English text, including biomedical documents since tokenization strategies can vary depending on language, task goals and other criteria. Previous approaches to biomedical tokenization lack guidance on how to modify existing tokenizers to new domains and how even to select them. Their idiosyncratic nature, detailed above, complicates this selection, modification and implementation (Barrett, 2012). Some authors also highlight the clear need for tokenization evaluation through the alignment and com-

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 40–49,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

parison of the results of different tokenizers (Habert et al., 1998). To address this challenge, this paper identifies and describes all the problematic cases that can be found when tokenizing a biomedical text. In addition, it includes a list of useful tokenizers and a comparison of their outputs on biomedical text samples.

The rest of the paper is organized as follows. Firstly, the most relevant related research is outlined. Secondly, the tokenizers are listed and their outputs are shown. The paper finishes with conclusions.

2 Related Work

Despite its importance, tokenization is often neglected in the literature (Dridan and Oepen, 2012). Most research has been focused on annotating corpus with token information (Ohta, et al., 2002; Tanabe et al., 2005; Verspoor, et al., 2012) and developing or adapting tokenizers to new domains (Tomanek et al., 2007; McClosky and Charniak, 2008). However, little attention has been paid to the analysis of the problematic cases that appear in the tokenization process and the different strategies used for the current available tokenization tools to solve them.

To the best of our knowledge, for the biomedical domain, there is only one work devoted to a comparison of several tokenizers (He and Kayaalp, 2006). In this study, He and Kayaalp made a first approximation of the challenging cases. As authors affirmed, it can be considered as a starting point since the limited scope of their effort prevented them from developing a more complete set of cases. Especially, the instances identified for biomedical named entities are insufficient. The study also includes a comparison of the output of 13 tokenizers on 78 biomedical abstracts from Medline, a corpus of biomedical literature compiled by the U.S. National Library of Medicine.

Due to the limitations in the categorization of the complex cases and the fact that many tokenization tools have been developed in recent years, this paper complete all these cases, update the list of tokenization tools and test them on a set of biomedical sentences, outlining the differences among tokenization schemes. This means, providing a qualitative guideline for the reader which aid the decision making process of choosing the right tokenizer. This decision will depend mainly on the downstream task. In addition, the critical issues identified, allow developers to know what should be taken into account when adapting or developing tokenization tools.

3 Material and Methods

3.1 Problematic cases

We could divide the potential complexities in the tokenization process into two major categories: those that apply across all domains and those that are more likely to be found in biomedical corpora, where there is a large amount of technical vocabulary (Clegg, 2008). All these difficulties, together with sentences extracted from the BioScope corpus (Vincze et al., 2008), in which authors such as Velldal et al. (2012) found problematic cases where tokenizers fail, are detailed below:

Common English complexities

- **Hyphenated compound words**

For example:

(1) *Normal chest **x-ray**.*

(2) ***2-year 2-month** old female with pneumonia.*

(3) *This may occur through the ability of **IL-10** to induce expression of the gene.*

- **Words with letters and slashes**

Slashes usually indicate alternatives (e.g. *differentiation/activation*) or measurement units (e.g. *ng/ml*). In addition, they often separate two or more entity references (e.g. *IL-12/CD34*). They may also denote the knock-out status of a certain gene with respect to an organism (e.g. *flt3L-/-mice*) (Tomanek et al., 2007). For example:

(4) *The maximal effect is observed at the IL-10 concentration of 20 **U/ml**.*

(5) *These results indicate that within the **TCR/CD3** signal transduction pathway both PKC and calcineurin are required for the effective activation of the IKK complex and NF-kappaB in T lymphocytes.*

- **Words with letters and apostrophes**

Apostrophes can indicate possessive (e.g. *years'*), words with single quotation (e.g. *'syntenic hits'*) and names (e.g. *O'Neill*). Examples of these might be the following:

(6) *The false positive rate of our predictor was estimated by the method of* ***D'Haeseleer*** *and Church 1855 and used to compare it to other prediction datasets.*

(7) *Small, scarred right kidney, below more than 2 standard deviations in size for* ***patient's*** *age.*

- **Words with letters and brackets**

There are basically four types of brackets: parentheses, square brackets, braces and angle brackets. For instance:

(8) *Of these, Diap1 has been most extensively characterized; it can block cell death caused by the ectopic expression of reaper, hid, and grim* ***(****reviewed in* ***[****26****])****.*

- **Abbreviations in capital letters and acronyms**

An abbreviation is a shortened form of a word or phrase. Usually, but not always, it consists of a letter or group of letters taken from the word or phrase. It must be taken into account in any tokenization process. An example of this may be the one shown below:

(9) *Mutants in Toll signaling pathway were obtained from* **Dr.** *S. Govind: cactE8, cactIIIG, and cactD13 mutations in the cact gene on Chromosome II.*

An acronym is an abbreviation formed from the initial components in a phrase or a word. These components may be individual letters (as in *SARS*; *severe acute respiratory syndrome*) or parts of words (as in *Ameslan*; *American Sign Language*).

Abbreviations and acronyms are commonly used in biomedical literature. For example, in the medical domain, writing favors brevity because time pressures often prevent medical specialists from describing clinical findings fully and abbreviations are a convenient way to shorten the sentences (Grange and Bloom, 2000).

Abbreviations and acronyms mainly refer to names, but abbreviations of adjectival expressions are often found in the biomedical domain (e.g. *CD8+* is an abbreviation of *CD8-positive*). For example:

(10) *The transcripts were detected in all the* ***CD4- CD8-****,* ***CD4+ CD8+****,* ***CD4+ CD8-****, and* ***CD4- CD8+*** *cell populations.*

- **Words with letters and periods**

Words with a period at the end usually indicate end of sentence. However, they may merely be abbreviations, such as *i.e.* and *e.g.* as shown in the following example:

(11) *Two stop codons of an iORF (****i.e****. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).*

- **Words with letters and numbers**

For example:

(12) *Selenocysteine and pyrrolysine are the* ***21st*** *and* ***22nd*** *amino acids, which are genetically encoded by stop codons.*

- **Words with numbers and one type of punctuation**

Some simple examples for numbers are: large numbers (e.g. *390,926*), fractions (e.g. *1/2*), percentages (e.g. *50%*), decimals (e.g. *0.001*) and ranges (e.g. *2-5*). These punctuation marks are: comma, forward slash, percent, period and en dash. Good illustrations extracted from the BioScope corpus are the following:

(13) *A total of* ***26,003*** *iORF satisfied the above criteria.*

(14) *The patient had prior x-ray on* ***1/2*** *which demonstrated no pneumonia.*

(15) *Indeed, it has been estimated recently that the current yeast and human protein interaction maps are only* ***50%*** *and* ***10%*** *complete, respectively 18.*

(16) *The dotted line indicates significance level* ***0.05*** *after a correction for multiple testing.*

(17) *E-selectin is induced within* ***1–2*** *h, peaks at* ***4–6*** *h, and gradually returns to basal level by 24 h.*

- **Numeration**

It is regarded as the act or process of counting or numbering. For instance:

(18) ***1.*** *Bioactivation of sulphamethoxazole (SMX) to chemically-reactive metabolites and subsequent protein conjugation is thought to be involved in SMX hypersensitivity.*

- **A hypertext markup symbol**

Some of the frequently observed hypertext markup symbols are *&lt;* and *&quot;* (for the double quotation mark). For example:

(19) *Bcd mRNA transcripts of* ***&lt;*** *or = 2.6 kb were selectively expressed in PBL and testis of healthy individuals.*

- **A URL**

An example would be the following:

(20) *Names of all available Trace Databases were taken from a list of databases at* ***http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml***

Biomedical English complexities

- **A DNA sequence**

For example:

(21) *Footprinting analysis revealed that the identical sequence* ***CCGAAACTGAAAAGG****, designated E6, was protected by nuclear extracts from B cells, T cells, or HeLa cells.*

- **Temporal expressions**

For instance:

(22) *This was last documented on the Nuclear Cystogram dated* ***1/2/01****.*

- **Chemical substances**

They include several symbols which may (or may not) denote word token boundary symbols such as parentheses, hyphens and slashes (Tomanek et al., 2007). Furthermore, chemical substances basically comprehend gene symbols, drug names and protein names, each of which has certain characteristics as described below.

Gene symbols

The names can indeed be divided into the following three categories (Proux et al., 1998).

– Names including special characters, i.e. upper cases, hyphen, digit, slash or brackets. For example, *Lam-B1* or *M(2)201*.
– Names in lower case and belonging to the general English language. For instance, *vamp* or *zip*.
– Names using lower case letters only without belonging to the language such as *zhr* or *sth*.

Drug names

In general, most drug names include:

– Particular letters from the chemical formula (e.g. T*ylenol*, which were generated from *n-aceryl-para-aminophenol*) as describe Gantner et al. (2002).
– Generic names such as *Thalomid*.
– Latin or Greek terminology.
– Parts or abbreviations of the company's name (e.g. *Baycol*, (*Bayer+colesterol*)).
– Low-frequency letters of the alphabet such as x or y (e.g. *x-trozine*).
– Acronyms like *Tigan* (that means *this is good against nausea*).

Protein names

Protein names can also be partitioned into three categories from their structure (Fukuda et al., 1998):

– Single words in upper case, numerical figures, and non-alphabetical letters which are mostly derived from gene name (e.g. *p53*).
– Compound words with upper case letters, numerical letters, and non-alphabetical letters. (e.g. *(IL-1)-responsive kinase*).
– Single word with only lower case letters (e.g. *insulin*).

Examples which appear in the BioScope corpus are the following:

(23) *These results reveal a central role for* ***CaMKIV/Gr*** *as a* ***Ca(2+)-regulated*** *activator of gene transcription in T lymphocytes.*

(24) *Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1* ***(IL-1)-responsive*** *cells blocked* ***IL-1-induced*** *gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.*

3.2 Tokenization strategies

The tools analyzed were the following: Freeling, Genia tagger, Gate Unicode tokenizer (GUT), JULIE LAB tokenizer (JLT), LingPipe, McClosky-Charniak parser (MCP), MedPost, NLTK tokenizer, OpenNLP tokenizer, Penn Bio tokenizer, Stanford POS tagger and Xerox tokenizer. Table 1 details all these tokenizers showing their references and websites.

These tools were tested on the set of examples extracted from the BioScope corpus listed in the previous section. Tables 2 to 24 detail the output from each tokenizer. Each row of the tables shows the list of tokenizers with the same output. The numbers of the tools refer to Table 1. In bold, decisions in which tokenizers do not match.

The outputs, for which there is no agreement among several tools and, therefore, correspond to a single tokenizer, are not shown in this paper due to the space limit. However, this information can be found in Supplementary Material.

Common English complexities

- **Hyphenated compound words**

Table 2: Tokenizers output for sentence (1)

Tokenizer	Output
1, 2, 3, 6, 8, 9, 10, 11	Normal‸chest‸**x-ray**‸.

Table 3: Tokenizers output for sentence (2)

Tokenizer	Output
1, 2, 6, 8, 9, 11, 12	**2-year‸2-month**‸old‸female‸with‸pneumonia‸.
3, 4, 5, 7	**2‸-‸year‸2‸-‸month**‸old‸female‸with‸pneumonia‸.

Table 4: Tokenizers output for sentence (3)

Tokenizer	Output
1, 2, 4, 6, 8, 9, 10, 11, 12	This‸may‸occur‸through‸the‸ability‸of‸**IL-10**‸to‸induce‸expression‸of‸the‸gene‸.‸
5, 7	This‸may‸occur‸through‸the‸ability‸of‸**IL‸-‸10**‸to‸induce‸expression‸of‸the‸gene‸.‸

- **Words with letters and slashes**

Table 5: Tokenizers output for sentence (4)

Tokenizer	Output
2, 6, 8, 9, 11, 12	The‸maximal‸effect‸is‸observed‸at‸the‸IL-10‸concentration‸of‸20‸**U/ml**‸.
3, 5, 7	The‸maximal‸effect‸is‸observed‸at‸the‸IL‸-‸10‸concentration‸of‸20‸**U‸/‸ml**‸.

Table 1: Overview of the 12 tools reviewed in the current study with their publications and website

	Tool	References	Website
1	Freeling	(Carreras, 2004; Padró and Stanilovsky, 2012)	http://nlp.lsi.upc.edu/freeling/
2	Genia	(Kulick et al., 2004; Tsuruoka et al., 2005; Tsuruoka and Tsujii, 2005)	http://www.nactem.ac.uk/tsujii/GENIA/tagger/
3	GUT	(Cunningham et al., 2002)	http://gate.ac.uk/sale/tao/splitch6.html#sec:annie:tokeniser
4	JLT	(Tomanek et al., 2007)	http://www.julielab.de/Resources/NLP+Tools.html
5	LingPipe	(Carpenter and Baldwin, 2011)	http://alias-i.com/lingpipe/
6	MCP	(McClosky and Charniak, 2008; McClosky, 2010)	http://nlp.stanford.edu/~mcclosky/biomedical.html
7	MedPost	(Smith et al., 2004)	ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz
8	NLTK	(Bird et al., 2009)	http://nltk.org/
9	OpenNLP	-	http://opennlp.apache.org/
10	Penn Bio	(Jin et al., 2006; McDonald and Pereira, 2005; McDonald et al., 2004)	http://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.html
11	Stanford	(Toutanova et al., 2003)	http://nlp.stanford.edu/software/tagger.shtml
12	Xerox	(Beesley and Karttunen, 2003)	http://open.xerox.com/Services/fst-nlp-tools/Consume/175

1, 4, 10	The‸maximal‸effect‸is‸observed‸at‸the‸IL-10‸concentration‸of‸20‸**U‸/‸ml.**

Table 6: Tokenizers output for sentence (5)

Tokenizer	Output
1, 2, 6, 8, 9, 11, 12	These‸results‸indicate‸that‸within‸the‸**TCR/CD3**‸signal‸transduction‸pathway‸both‸PKC‸and‸calcineurin‸are‸required‸for‸the‸effective‸activation‸of‸the‸IKK‸complex‸and‸NF-kappaB‸in‸T‸lymphocytes‸.
3, 4, 5, 7, 10	These‸results‸indicate‸that‸within‸the‸**TCR‸/‸CD3**‸signal‸transduction‸pathway‸both‸PKC‸and‸calcineurin‸are‸required‸for‸the‸effective‸activation‸of‸the‸IKK‸complex‸and‸NF‸-‸kappaB‸in‸T‸lymphocytes‸.

- **Words with letters and apostrophes**

Table 7: Tokenizers output for sentence (6)

Tokenizer	Output
1, 2, 4, 8, 9, 10, 11, 12	The‸false‸positive‸rate‸of‸our‸predictor‸was‸estimated‸by‸the‸method‸of‸**D'Haeseleer**‸and‸Church‸1855‸and‸used‸to‸compare‸it‸to‸other‸prediction‸datasets‸.
3, 5, 6, 7	The‸false‸positive‸rate‸of‸our‸predictor‸was‸estimated‸by‸the‸method‸of‸**D‸'‸Haeseleer**‸and‸Church‸1855‸and‸used‸to‸compare‸it‸to‸other‸prediction‸datasets‸.

Table 8: Tokenizers output for sentence (7)

Tokenizer	Output
1, 2, 4, 6, 8, 9, 10, 11, 12	Small‸,‸scarred‸right‸kidney‸,‸below‸more‸than‸2‸standard‸deviations‸in‸size‸for‸**patient‸'s‸age**‸.
3, 5, 7	Small‸,‸scarred‸right‸kidney‸,‸below‸more‸than‸2‸standard‸deviations‸in‸size‸for‸**patient‸'‸s**‸age‸.

- **Words with letters and brackets**

Table 9: Tokenizers output for sentence (8)

Tokenizer	Output
1, 2, 5, 7, 8, 11, 12	Of‸these‸,‸Diap1‸has‸been‸most‸extensively‸characterized‸;‸it‸can‸block‸cell‸death‸caused‸by‸the‸ectopic‸expression‸of‸reaper‸,‸hid‸,‸and‸grim‸**(**‸reviewed‸in‸**[‸26‸]‸)**‸.

- **Abbreviations in capital letters and acronyms**

Table 10: Tokenizers output for sentence (9)

Tokenizer	Output
4, 6, 8, 11	Mutants‸in‸Toll‸signaling‸pathway‸were‸obtained‸from‸**Dr.**‸S.‸Govind‸:‸cactE8‸,‸cactIIIG‸,‸and‸cactD13‸mutations‸in‸the‸cact‸gene‸on‸Chromosome‸II‸.
2, 5, 7	Mutants‸in‸Toll‸signaling‸pathway‸were‸obtained‸from‸**Dr‸.‸S‸.**‸Govind‸:‸cactE8‸,‸cactIIIG‸,‸and‸cactD13‸mutations‸in‸the‸cact‸gene‸on‸Chromosome‸II‸.

Table 11: Tokenizers output for sentence (10)

Tokenizer	Output
2, 6, 8, 9, 12	The‸transcripts‸were‸detected‸in‸all‸the‸**CD4-‸CD8-**‸,‸**CD4+‸CD8+**‸,‸**CD4+‸CD8-**‸,‸and‸**CD4-‸CD8+**‸cell‸populations‸.
1, 3, 4, 7, 10, 11	The‸transcripts‸were‸detected‸in‸all‸the‸**CD4‸-‸CD8‸-**‸,‸**CD4‸+‸CD8‸+**‸,‸**CD4‸+‸CD8‸-**‸,‸and‸**CD4‸-‸CD8‸+**‸cell‸populations‸.

- **Words with letters and periods**

Table 12: Tokenizers output for sentence (11)

Tokenizer	Output
1, 6, 11, 12	Two‸stop‸codons‸of‸an‸iORF‸(‸**i.e.**‸the‸inframe‸and‸C-terminal‸stops‸)‸can‸be‸any‸combination‸of‸canonical‸stop‸codons‸(‸TAA‸,‸TAG‸,‸TGA‸)‸.
2, 8	Two‸stop‸codons‸of‸an‸iORF‸(‸**i.e‸.**‸the‸inframe‸and‸C-terminal‸stops‸)‸can‸be‸any‸combination‸of‸canonical‸stop‸codons‸(‸TAA‸,‸TAG‸,‸TGA‸)‸.
4, 7	Two‸stop‸codons‸of‸an‸iORF‸(‸**i‸.‸e‸.**‸the‸inframe‸and‸**C‸-‸terminal**‸stops‸)‸can‸be‸any‸combination‸of‸canonical‸stop‸codons‸(‸TAA‸,‸TAG‸,‸TGA‸)‸.

- **Words with letters and numbers**

Table 13: Tokenizers output for sentence (12)

Tokenizer	Output
1, 2, 4, 5, 6, 7, 8, 9, 11, 12	Selenocysteine‸and‸pyrrolysine‸are‸the‸**21st**‸and‸**22nd**‸amino‸acids‸,‸which‸are‸genetically‸encoded‸by‸stop‸ codons‸.

- **Words with numbers and one type of punctuation**

Table 14: Tokenizers output for sentence (13)

Tokenizer	Output
1, 5, 6, 8, 9, 10, 11, 12	A‸total‸of‸**26,003**‸iORF‸satisfied‸the‸above‸criteria‸.
2, 3, 4, 7	A‸total‸of‸**26**‸**,**‸**003**‸iORF‸satisfied‸the‸above‸criteria‸.

Table 15: Tokenizers output for sentence (14)

Tokenizer	Output
1, 2, 6, 8, 9, 11, 12	The‸patient‸had‸prior‸**x-ray**‸on‸**1/2**‸which‸demonstrated‸no‸pneumonia‸.
4, 5, 7	The‸patient‸had‸prior‸**x**‸**-**‸**ray**‸on‸**1**‸**/**‸**2**‸which‸demonstrated‸no‸pneumonia‸.
3, 10	The‸patient‸had‸prior‸**x-ray**‸on‸**1**‸**/**‸**2**‸which‸demonstrated‸no‸pneumonia‸.

Table 16: Tokenizers output for sentence (15)

Tokenizer	Output
3, 4, 5, 6, 7, 8, 9, 10, 11	Indeed‸,‸it‸has‸been‸estimated‸recently‸that‸the‸current‸yeast‸and‸human‸protein‸interaction‸maps‸are‸only‸**50**‸**%**‸and‸**10**‸**%**‸complete‸,‸respectively‸18‸.

Table 17: Tokenizers output for sentence (16)

Tokenizer	Output
1, 2, 4, 5, 6, 8, 9, 10, 11, 12	The‸dotted‸line‸indicates‸significance‸level‸**0.05**‸after‸a‸correction‸for‸multiple‸ testing‸.
3, 7	The‸dotted‸line‸indicates‸significance‸level‸**0**‸**.**‸**05**‸after‸a‸correction‸for‸multiple‸ testing‸.

Table 18: Tokenizers output for sentence (17)

Tokenizer	Output
1, 2, 8, 9, 10, 11, 12	E-selectin‸is‸induced‸within‸**1–2**‸**h**‸,‸peaks‸at‸**4–6**‸**h**‸,‸and‸gradually‸returns‸to‸basal‸level‸by‸**24**‸**h**‸.
4, 7	E-selectin‸is‸induced‸within‸**1**‸**–**‸**2**‸**h**‸,‸peaks‸at‸**4**‸**–**‸**6**‸**h**‸,‸and‸gradually‸returns‸to‸basal‸level‸by‸ **24**‸**h**‸.

- **Numeration**

Table 19: Tokenizers output for sentence (18)

Tokenizer	Output
1, 2, 3, 5, 7, 8, 9, 10, 11, 12	**1**‸**.**‸Bioactivation‸of‸sulphamethoxazole‸(‸SMX‸)‸to‸chemically-reactive‸metabolites‸and‸subsequent‸protein‸conjugation‸is‸thought‸to‸be‸involved‸in‸SMX‸hypersensitivity‸.
4, 6	**1.**‸Bioactivation‸of‸sulphamethoxazole‸(‸SMX‸)‸to‸chemically-reactive‸metabolites‸and‸subsequent‸protein‸conjugation‸is‸thought‸to‸be‸involved‸in‸SMX‸hypersensitivity‸.

- **A hypertext markup symbol**

Table 20: Tokenizers output for sentence (19)

Tokenizer	Output
2, 4, 5, 8	Bcd‸mRNA‸transcripts‸of‸**&**‸**lt**‸**;**‸or‸=‸**2.6**‸**kb**‸were‸selectively‸expressed‸in‸PBL‸and‸testis‸of‸healthy‸individuals‸.
9, 12	Bcd‸mRNA‸transcripts‸of‸**&lt**‸**;**‸or‸=‸**2.6**‸**kb**‸were‸selectively‸expressed‸in‸PBL‸and‸testis‸of‸healthy‸individuals‸.
3, 7	Bcd‸mRNA‸transcripts‸of‸**&**‸**lt**‸**;**‸or‸=‸**2**‸**.**‸**6**‸**kb**‸were‸selectively‸expressed‸in‸PBL‸and‸testis‸of‸healthy‸individuals‸.

- **A URL**

Table 21: Tokenizers output for sentence (20)

Tokenizer	Output
2, 6, 8	Names‸of‸all‸available‸Trace‸Databases‸were‸taken‸from‸a‸list‸of‸databases‸at‸**http**‸**:**‸**//www.ncbi.nlm.nih.gov/blast/mmtrace.shtml**
3, 5, 7	Names‸of‸all‸available‸Trace‸Databases‸were‸taken‸from‸a‸list‸of‸databases‸at‸**http**‸**:**‸**/**‸**/**‸**www**‸**.**‸**ncbi**‸**.**‸**nlm**‸**.**‸**nih**‸**.**‸**gov**‸**/**‸**blast**‸**/**‸**mmtrace**‸**.**‸**shtml**
11, 12	Names‸of‸all‸available‸Trace‸Databases‸were‸taken‸from‸a‸list‸of‸databases‸at‸**http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml**

Biomedical English complexities

- **A DNA sequence**

Table 22: Tokenizers output for sentence (21)

Tokenizer	Output

1, 2, 4, 5, 6, 7, 8, 9, 11, 12	Footprinting‸analysis‸revealed‸that‸the‸identical‸sequence‸**CCGAAACTGAAAAGG**‸,‸designated‸E6‸,‸was‸protected‸by‸nuclear‸extracts‸from‸B‸cells‸,‸T‸cells‸,‸or‸HeLa‸cells‸.

- **Temporal expressions**

Table 23: Tokenizers output for sentence (22)

Tokenizer	Output
2, 6, 8, 9, 11, 12	This‸was‸last‸documented‸on‸the‸Nuclearv‸Cystogram‸dated‸**1/2/01**‸.
1, 3, 4, 7, 10	This‸was‸last‸documented‸on‸the‸Nuclearv‸Cystogram‸dated‸**1‸/2‸/‸01**‸.

- **Chemical substances**

Table 24: Tokenizers output for sentence (23)

Tokenizer	Output
6, 8	These‸results‸reveal‸a‸central‸role‸for‸**CaMKIV/Gr**‸as‸a‸**Ca‸(2+‸)‸-regulated**‸activator‸of‸gene‸transcription‸in‸T‸lymphocytes‸.
1, 3, 4, 7	These‸results‸reveal‸a‸central‸role‸for‸**CaMKIV‸/‸Gr**‸as‸a‸**Ca‸(‸2+‸)‸-‸regulated**‸activator‸of‸gene‸transcription‸in‸T‸lymphocytes‸.

Table 25: Tokenizers output for sentence (24)

Tokenizer	Output
1, 2, 6, 8, 11	Expression‸of‸a‸highly‸specific‸protein‸inhibitor‸for‸cyclic‸**AMP-dependent**‸protein‸kinases‸in‸**interleukin-1‸(‸IL-1‸)‸-‸responsive**‸cells‸blocked‸**IL-1-induced**‸gene‸transcription‸that‸was‸driven‸by‸the‸kappa‸immunoglobulin‸enhancer‸or‸the‸human‸immunodeficiency‸virus‸long‸terminal‸repeat‸.

4 Conclusions

This paper analyzed the problematic cases that can be found when tokenizing a biomedical text. In addition, it listed a set of potentially useful tokenizers and tested them on biomedical sentences.

Identifying the complex cases that introduce this domain and knowing what types of behavior are expected from available tokenizers in each of these cases is vital. This will enable researchers to be aware of those aspects which are especially challenging when developing new tools or adapting existing ones. In addition, it will aid the process of selecting the right tokenizer according to the most appropriate tokenization scheme for the downstream application. This will facilitate to lose the minimum of information. Obviously, other factors like technical, usability of functional criteria should be taken into account in such decision.

The experiments carried out showed a widely variation on the results. This variability was expected since there is no a single tokenization method. Neither of the tools produced identical output. Tokenizers pair that coincided in the same strategy or scheme in over 75% of cases were Genia tagger and NLTK tokenizer as well as Stanford POS tagger and NLTK tokenizer.

Regarding the challenging problems where there was more disagreement (less than 35% agreement) and, therefore, presented more difficulties for the tokenization tools are, the hypertext markup symbol, URLs and chemical substances. The latter was assumed since biomedical terminology is currently one of the most challenging research topics in NLP.

Among the cases with more than 80% agreement, it can be found: hyphenated compound words, words with letters and numbers, words with numbers and one type of punctuation and DNA sequences.

References

Andrew B. Clegg. 2008. Computational-Linguistic Approaches to Biological Text Mining. PhD thesis. School of Crystallography Birkbeck, University of London.

Benoit Habert, Gilles Adda, M. Adda-Decker, P. Boula de Marëuil, S. Ferrari, O. Ferret, G. Illouz, and P. Paroubek. 1998. Towards tokenization evaluation. In *Proceedings of the 1st International Language Resources and Evaluation*, p. 427-431. Granada, Spain.

Bob Carpenter, and Breck Baldwin. 2011. Natural Language Processing with LingPipe 4. LingPipe Publishing, New York.

Bob Grange, D.A. Bloom. 2000. Acronyms, abbreviations and initialisms. *BJU international*, vol. 86, no 1, p. 1-6.

Dan Jurafsky and James H. Martin. 2009. *Speech and Language Processing, 2nd Edition*. NJ: Pearson.

David Campbell and Stephen Johnson. 2001. Comparing syntactic complexity in medical and non-medical corpora. In *Proceedings of the AMIA Symposium*, p. 90. American Medical Informatics Association.

David McClosky, Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of the Association for Computational Linguistics*. Columbus, Ohio.

David McClosky. 2010. Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing. PhD thesis. Department of Computer Science, Brown University.

Denys Proux, Francois Rechenmann, Laurent Julliard, Violaine Pillet, and Bernard Jacq. 1998. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome informatics series*, 72-80.

Erik Velldal, Lilja Øvrelid, Jonathon Read and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational linguistics*, 38:2, 369-410.

Florian Gantner, Christian Schweiger, and Michael Schlander. 2002. Naming, classification, and trademark selection: implications for market success of pharmaceutical products. *Drug information journal*, 36:807–824.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 168-175. Portland, OR, USA.

Jonathan Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proceedings of the 14th Conference on Computational Linguistics-Volume 4*, 1106-1110.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, p. 49-57. Portland, OR, USA.

Karin Verspoor, et al. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics* 13.1 (2012): 207.

Kenneth R. Beesley and Lauri Karttunen. 2003. Finite state morphology. *Stanford: CSLI publications*.

K. I. Fukuda, T. Tsunoda, A. Tamura, and T Takagi. 1998. Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput*, 707-18.

Kristina Toutanova, Dan Klein, Chritopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, p. 173-180. Portland, OR, USA.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(suppl 1):S3.

Lluís Padró, and Evgeny Stanilovsky. 2012. Freeling 3.0: towards wider multilinguality. In *Proceedings of the 8th International Language Resources and Evaluation*. Istanbul, Turkey.

L. Smith, T. Rindflesch, and W.J. Wilbur. 2004. MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics, 20:14, 2320-2321.

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37:6, 512-526.

Neil Barrett. 2012. Natural language processing techniques for the purpose of sentinel event information extraction. PhD thesis. University of Victoria, Canada.

Rebecca Dridan and Stephan Oepen. 2012. Tokenization: Returning to a Long Solved Problem. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, p.

378-382. Association for Computational Linguistics, Jeju, Korea.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Ryan McDonald, and Fernando Pereira. 2005. Identifying gene and protein mentions in text using conditional random fields. *BMC bioinformatics*, 6:S1, S6.

Ryan T. McDonald, R. Scott Winters, Mark Mandel, Yang Jin, Peter S. White, and Fernando Pereira. 2004. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 20:17, 3249-3251.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. 2004. Integrated Annotation for Biomedical Information Extraction. *NAACL/HLT Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, 61-68. Boston.

Steven Bird, Ewan Klein, Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference*, 73-77. San Diego, California.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11), S9.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Language Resources and Evaluation,*. Lisbon, Portugal.

Yang Jin, Ryan T McDonald, Kevin Lerman, Mark A. Mandel, Steven Carroll, Mark Y. Liberman, Fernando C. Pereira, Raymond S. Winters, and Peter S White. 2006. Automated recognition of malignancy mentions in biomedical literature. *BMC bioinformatics*, 7:1, 492.

Ying He and Mehmet Kayaalp. 2006. A Comparison of 13 Tokenizers on MEDLINE. The Lister Hill National Center for Biomedical Communications, Tech. Rep. LHNCBC-TR-2006-003.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, Jun’ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics - 10th Panhellenic Conference on Informatics*. Springer Berlin Heidelberg, p. 382-392.

Yoshimasa Tsuruoka, and Jun’ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 467-474. Portland, OR, USA.

Zellig S. Harris. 2002. The structure of science information. *Journal of biomedical informatics*, 35:4, 215-221.