Natural Language Processing

A General Discussion

Table of Contents

1. Discussion of some of the challenges presented by Twitter and other social media as well as NLP tools such as tokenisers, POS taggers, chunkers and parsers.

Both social media platforms and microblogging services platforms have grown massively in the past 6 years. The former are software tools enabling connected users to share ideas and opinions [1]. Microblogging services, such as Twitter, are a bit more restrictive. They allow the broadcasting of short posts (usually between 140-200 characters). As shown in [2], the number of monthly active users has increased by 30 million per month on average from 2010 to 2016. This has naturally led to a rise of interest for research topics in text mining, sentiment analysis and opinion mining. This new way of written informal communication offers several major challenges for text processing tools such as *tokenisers*, *part-of-speech taggers* (a.k.a. *POS* taggers), *chunkers* and *parsers*. These terms are briefly described as follows:

i) A *segmentation/tokenising* task aims at dividing a string sequence into bags of related element types (such as words, symbols or other element types), based on a defined instance of sequence of characters; namely the *token* [3].
ii) The *POS* task is responsible for labelling words against defined lexical categories.
iii) *Chunking* is the mechanism of "identifying non-overlapping linguistic groups" [4] (such as noun phrase, propositional phrase, verb phrase, etc.).
iv) Parsers are used to generate syntax structures which represent possible trees for a sequence of tokens [5].

Originally these tools were designed to analyse text coming from documents such as reports, pieces of literature, editorial articles, where there is intrinsic assumption relating to the usage of the English language. Social media in general and specifically micro bloggings disenfranchise themselves from such constraints. Consequently, new ways of expressing feelings and opinions have emerged, away from the conventional grammar rules.

These tools are limited in their capacity to perform their job adequately, for the following reasons:

i) Social media are open-domain [7], there is nothing that prevent an author to write a message which does not relate to the context of the domain under discussion (e.g. *Tab tweets /Row84* "Trump: win by electoral college". This sentence can be classified as off-topic, as there is no direct relation between "Brexit" and the American election result. This means that applying such tools for an off-topic sentence is irrelevant, time consuming and computing resource wasting.
ii) An author could write in a language that is not supported but the tokeniser/chunker or parser (e.g. *Tab tweets /Row7* "No iba a salir el Brexit y salió."*). Usually, tokenisesr, such as the NLTK tokeniser, require the expected language under analysis to be passed as a parameter. When 'English' is set, then the tokeniser cannot break up this sentence.*
iii) Data is usually sparse [7], it usually contains many abbreviated words (due to the word limitation). An example of an abbreviation is "inc" *Tab tweets /Row40.*
iv) Tweets are not edited. They could contain incorrect grammar structures or spelling mistakes [9]. A spelling mistake example is "skyslide" (e.g. *Tab tweets /Row4)*. It should have a capital letter as this is not a noun but a proper noun (the name of the tallest building in New York).
v) Even an improved *POS,* such as the *Tweet NLP* [8]*,* is limited by the lack of context [6], or the usage of sarcasm. *Tab tweets /Row28* shows: "Solidarity means solidarity (like Brexit means Brexit) - but what we do mean by solidarity? Empathy? Gratitude? Or… ". The nouns "solidarity", "empathy" and "gratitude" are usually attributed to a positive sentiment, however because of the context of this sentence ("what do we mean by"), the author most probably means these concepts are absent following Brexit. Therefore, implying a negative sentiment. It also fails to identify "obscure symbols and artefacts of tokenization errors" [9]. There are no such examples in the "Brexit" file.

vi) The *segmentation* and *tokenisation* face another set of challenges relating to emoticons, email addresses, URL or hashtags. Examples of hashtags and URL can be found in the "Brexit" mini-corpus tabs *hashtags* and *urls.* They usually carry a meaning and they need to be labelled appropriately. Although tokenisers could succeed in tagging such elements, the *POS/chunkers* would not be able to classify them correctly.

vii) The aforementioned tools do also ignore the content of URL referenced in a tweet or retweets, therefore important meaning could be lost. There is an example of a URL in *Tab tweets /Row48* "(…*) Supreme Court rules* https://t.co/dFwm6RGPsH*".* Retweets can be found in the tab *retweets.*

The issues relating to the handling of hyphenated words (e.g. "fine-tuning"), white space (i.e. " "), English enclitics (e.g. I'm), numerical and special expressions such as (date, time, price, citations, etc.) have not been studied here, as these issues are not specific to social media/microblogging engines.
Because of these intrinsic limitations, statistical methods, such as machine learning algorithm (e.g. Naive Bayes, Maximum Entropy, Support Vector Machines, etc.) are often used to complement and refine the text mining to extract contextual meaning/sentiment.

References

[1] *English Oxford Living Dictionary*. [Online] Available at: https://en.oxforddictionaries.com/definition/social_media [Accessed: 08-December-2016].

[2] *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2016 (in millions)* (2016) [Online], Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [Accessed: 05-December-2016].

[3] Bird S, Klein E, Loper E. (2009), *Natural Language Processing with Python*, O'Reilly Media, Inc, Sebastopol, Section pp102-107, p179

[4] Bird S., Loper E. (2016) *Source code for nltk.chunk* [Online]. Available At: http://www.nltk.org/_modules/nltk/chunk.html [Accessed: 08-December-2016].

[5] *nltk.parse package* (2016). Available At: http://www.nltk.org/api/nltk.parse.html [Accessed: 04-December-2016]

[6] Ramanujam S. (2014), *Twitter NLP Example: How to Scale Part-of-Speech Tagging with MPP (Part 1)*. Available At: from https://blog.pivotal.io/data-science-pivotal/products/twitter-nlp-example-how-to-scale-part-of-speech-tagging-with-mpp-part-1 [Accessed: 12-December-2016]

[7] Saif H. (2012), *Sentiment Analysis of Microblogs*, Mining the New World, Technical Report KMI-12-2, Open University

[8] Tweet NLP [Online], Available At: http://www.cs.cmu.edu/~ark/TweetNLP/#pos [Accessed: 08-December-2016]

[9] Gimpel K., Schneider N., O'Connor B, Das D., Mills D., Eisenstein J., Heilman M., Yogatama D., Flanigan J., Smith N.A. (NA), *Part-of-SpeechTaggingforTwitter: Annotation,Features,andExperiments* [Online], Available At: http://www.cs.cmu.edu/~ark/TweetNLP/gimpel+etal.acl11.pdf, Accessed [13-December-2016]

2. Discussion relative to the use of mathematical models in Linguistics

Part I – Noam Chomsky considers that statistical models, and by extension empirical sciences, are not fit for providing insight into the theory of languages. Artificial intelligence (AI) specialises in collecting data relating to language patterns, therefore it is successful in describing 'what' happens and 'how', but it cannot explain 'why' a linguistic construction happens in the first instance [4]. Chomsky famously said "you can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles."
Furthermore, Chomsky considers that the successes achieved by engineering techniques satisfy themselves with the approximation of the *truth*. This happens with machine learning where models are considered valid when they come close enough to the goal [1]. For Chomsky throwing more data to the algorithm improves the approximation, but does not provide more insight on the internal mechanism that generates the original behaviour.
Chomsky dislikes statistical models for four main technical reasons:

- Chomsky stresses that probabilistic models cannot model the language, as they are not the model used by human brain as a tool to decipher meaning in sentences [3].
- They are complex by nature. Chomsky argues that linguistics should be explained in simple and comprehensible terms, and that the use of complex AI models coupled with parameter tuning is the not right tool for explaining the theory of language [3].
- In Syntactic Structures [2], Noam Chomsky goes a step further in the analysis of nonsensical but grammatically correct sentences. For example, the sentence: 'colourless green ideas sleep' would be discarded by AI algorithms that attempt to search for a "high order of statistical approximation to English", although it is grammatically correct. Indeed, the sentence would be considered as 'remote' from English, as no such sentence has been written in any text before. Therefore, it would be considered as an outlier and discarded from the analysis.
- Chomsky also emphasises that so far, no AI machine has been able to learn a language [3]. Only humans are capable of such a task. It is the result of evolution that has shaped the inner working of our brain (which is yet far from being fully understood by scientists). Therefore, the usage of statistical models that mimics this behaviour should be discarded.

References

[1] Chomsky N. (NA). *Q&A transcript* [Online], http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html. The Golden Age — A Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience [Accessed: 08-December-2016].

[2] Chomsky N. (2002). *Syntactic Structures* [Online], https://archive.org/stream/NoamChomskySyntcaticStructures/Noam%20Chomsky%20-%20Syntcatic%20structures_djvu.txt [Online]. [Accessed: 08-December-2016].

[3] Norvig P. (NA). *On Chomsky and the Two Cultures of Statistical Learning* [Online]. Brains, Minds, and Machines symposium held during MIT's. Available at: http://norvig.com/chomsky.html [Accessed: 08-December-2016].

[4] Katz Y., (2012), *Noam Chomsky on forgotten methodologies in artificial intelligence* [Online], Available at: https://www.youtube.com/watch?v=yyTx6a7VBjg [Accessed: 08-December-2016]

Part II – Engineering, mathematical and probabilistic techniques have evolved very rapidly in the last 60 years. These developments coupled with the spread of accessibility of powerful computers and the ability to process large volume of data have transformed the linguistics analysis landscape. The use of these techniques, in the context of empirical sciences, have given birth to numerous technologies that are used daily by millions of users around the world such as search engines, speech recognition and machine translation [1]. At a more granular level, they have also proven their success in the fields of language acquisition, language change and variation as well as language generation and comprehension [2].

The probabilistic approach for linguistics analysis is legitimate as it offers "(…) a systematic study of the structure and behaviour of the physical and natural world through observation and experiment" [2][3]. Probabilistic models usually offer good prediction capability and, even when they contain large number of tuning parameters, they provide insight through the analysis of the successes and failures they generate [2]. The analysis of language is not different from the analysis of other physical phenomena. It is highly stochastic and therefore, although probabilistic methods are not perfect, they are the best tools available. The human brain (through the very complex neuron mesh) could be using probabilistic decision making process, without the humans' awareness of it [2].
Probabilistic models have achieved numerous successes in the areas mentioned above, but also in coreference resolution, part of speech tagging and parsing [2]. Probabilistic models coupled with statistical techniques are well suited for representing and interpreting linguistic facts. For example, statically-trained finite- state machines can recognise grammatical and ungrammatical constructions as well as classifying their frequency of appearance [2].
Furthermore, probabilistic methods offer the most promising developments in understanding how the human brain manage the construction and exchange of a language. These methods have proven their efficiency in the fields of language acquisition (i.e. the learning of a new language), language variation (i.e. dialectology and typology) and language change (i.e. change in the relative frequency construction) [5]. Abney argues that the phenomenon of language acquisition, from childhood, is a trial and error process, where the brain possibly assigns probability against potential constructions. Consequently, it is likely that grammar follows a stochastic process.
Stochastic models are also paramount in the analysis of language generation and comprehension, as well as finding the presence of variation and noise [5]. Abney argues that the current nomenclature in linguistics is concerned by the "well-formedness of sentences for which the intended structure is specified", a.k.a. structure judgements. However, it does not focus on the "performance" data, i.e. the production and comprehension of the language. *Performance* data is usually regarded as a domain for psychologists, not linguists. For Abney, the *performance data* contain information of linguistic importance and relevance. At first, he opposes the concepts of "grammatically and ambiguity judgements about sentences as opposed to structures". These judgements cannot easily be accounted for by usual grammars due to their stochastic or weighted nature. One aspect is the rare usage of common words that may change the meaning of a sentence (e.g. *are* could be used as a conjugated verb or as a noun indicating 100 hectares). When the sentence construction allows for a potential ambiguity in meaning, the longer the sentence, the higher the chance is for this issue to appear or compound itself. Therefore, a grammatically correct sentence could have an absurd meaning. Due to potential gap between grammatical and ambiguity in structural prediction and human perception, it is necessary to use weighted grammar to identify the more likely correct structure from the universe of possible structures.
However, there are intrinsic limitation of weighted grammars to explain human "natural-sounding sentences" [5]. This is where computational distributional and statistics methods can help in:

i) Computing the optimum parse that can disambiguate meaning.
ii) Defining the degree of goodness for speech production and comprehension. This mixes ranking the grammar structures from best to worst grammatical or ungrammatical structures with degree of naturalness and structural preference (e.g. a "longest-match" preference). The naturalness represents defined collocation (i.e. "how you say it") and

selection restrictions (i.e. how a sentence or construction is more or less plausible compared to another).

iii) Supporting error tolerance and correction, as applied by humans. A sentence could contain a grammatical error (e.g. we goes to the cinema) but the intended grammatical structure is obvious and can be corrected using statistical techniques.

iv) Learning on the fly, for example learning a new sub-categorisation property of a verb (from verb category to transitive or intransitive verb categories). This enables adaption in the presence of noise and variance in a linguistics context.

v) Supporting the acquisition of grammars and lexica to improve the chance of correct parsing.

There are numerous objections that can be raised against the use of statistical and probabilistic methods. First, as argued by Chomsky [1], the empirical nature of these approaches make them unfit for use in a theoretical framework. As mentioned in the first section, although the approach taken is indeed based on trial and error, it does follow a scientific setting; experiences are robust and repeatable. Like in the field of physics, the analysis of the observed behaviour enables the computational linguist to draw rules from observed behaviours. These rules may be limited and imperfect at first. With time, they get confirmed (or rejected), refined or even superseded.

A corollary criticism lies in the fact that this engineering approach focuses on linguistics behaviour observations, rather than a deep understanding of the theory of linguistics. In other words, without the theoretical knowledge, drawing conclusion from observations produce a weak form of knowledge. By extension the use of statistical tools for linguistics analysis should be discarded. This is a strong statement and it contradicts the inductive learning principles in artificial intelligence. Once a prediction is made, following an induction, the scientist tests the performance of the prediction against an independent dataset. This is the responsibility of the scientists to establish whether the performance levels make the model acceptable or not. From the model emerges a new form of knowledge that can be compared and/or compounded to existing knowledge. Furthermore, in the case of neural networks for example, neural weights change with new data. Therefore, the models are adaptable and can still be back tested to check their continuous relevance.

Second, Markov models (and by extension statistical models) are unfit for mimicking language acquisition as they require tuning a very large number of parameters [1]. As describe in [2], this is a narrow view of statistical model, which only focuses on a finite-state automat. More modern, stochastic based models have been developed, such as BLOSUM [4]. They can detect rhymes which are undetectable by standard statistical models, using probabilistic block substitution matrix coupled with a very small set of parameters.

Third, Chomsky [1] asserts that only humans can learn a language. Technology should not try to replicate this using statistical model. Chomsky considers this is a waste of time. So far technology can only provide limited success in terms of language acquisition and performance. As mentioned in [6], IBM Watson is not able to “understand the difference between polite and offensive speech”. Science and technologies do not claim that probabilistic methods are a true representation of how human speech works. However, this is the best tool so far. Even if this is not perfect, scientists should not discard research in these areas, as they represent other potential avenues that can lead, with time, to the understanding of how language is produced, comprehended, acquired and how it varies.

Chomsky [7] also argues that the engineering techniques satisfy themselves with the approximation of the *truth*. Science is only interested by the *truth,* not an approximation of it. Stochastic queue models approximate information when this one is not available or difficult to gather. However, they can provide, in certain situations, more insight than deterministic model. The random nature of stochastic models enables the analysis of the consequences of small variations in the inputs, which is not possible in pure deterministic models.

There are clearly two cultures in linguistics which are diametrically opposed in terms of i) the definition of the aim of linguistics analysis and ii) the way to achieve this aim. The pure linguistics approach led by

Chomsky imposes a research of the truth in Platonic terms. In other words, only the deep understanding of linguistics phenomena can direct research in using the data required to confirm (or infirm) the proposed theorem. The success is in this area is measured in terms of purity of the understanding. This is a valid argument. However, it should not overshadow empirical research that has so far, although still limited, proved to produce numerous visible successes in terms of speech recognition and robot to human interaction. It has also led to discovery of new relationships and ideas hidden behind the concept of performance that were ignored by traditional linguistics; such as disambiguation, degrees of grammaticality, naturalness, error tolerance and learning on fly. They are complex and intransigent linguistics problems that require modern mathematical techniques.

References

[1] Chomsky N., (NA). *Q&A transcript* [Online], http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html [Online]. The Golden Age — A Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience [Accessed: 10-December-2016].

[2] Norvig P., (NA). *On Chomsky and the Two Cultures of Statistical Learning* [Online]. Brains, Minds, and Machines symposium held during MIT's. Available at: http://norvig.com/chomsky.html [Accessed: 08-December-2016].

[3] *English Oxford Living Dictionary* [Online]. Available at: https://en.oxforddictionaries.com/definition/science [Accessed: 05-December-2016].

[4] Hirjee H., Brown D. (2013). *"Using Automated Rhyme Detection to Characterize Rhyming Style in Rap Music*", Empirical Musicology Review, Vol. 5, No. 4

[5] Steve A., (1996), *Statistical Methods and Linguistics*, in Klavans and Resnik (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press.

[6] Steadman I., (2013). *IBM's Watson is better at diagnosing cancer than human doctors* [Online]. Wired Artificial Intelligence. Available at: http://www.wired.co.uk/article/ibm-watson-medical-doctor [Accessed: 07-December-2016].

[7] Katz Y., (2012), *Noam Chomsky on forgotten methodologies in artificial intelligence* [Online], Available at: https://www.youtube.com/watch?v=yyTx6a7VBjg [Accessed: 08-December-2016]
http://www.businessinsider.com/phrases-only-wall-street-understands-2014-1?IR=T

[11] Hutto, C.J., Gilbert, E., (NA), *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text* [Online], Available at: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf [Accessed: 19-Jan-2017]