Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 212 (2015) 187 – 191

Procedia
Social and Behavioral Sciences

MULTIMODAL COMMUNICATION IN THE 21ST CENTURY: PROFESSIONAL AND ACADEMIC CHALLENGES. 33rd Conference of the Spanish Association of Applied Linguistics (AESLA), XXXIII AESLA CONFERENCE, 16-18 April 2015, Madrid, Spain

Fuzzy Grammaticality Models: A Tool for Web Language Analysis

M. Dolores Jiménez-López*, Adrià Torrens Urrutia

Universitat Rovira i Virgili, Av. Catalunya 35, Tarragona 43002, Spain

Abstract

In this paper, we highlight the need to propose formal models that consider grammaticality as a gradient property instead of the categorical view of grammaticality defended in theoretical linguistics. Given that deviations from the norm are inherent to the spontaneous use of language, linguistic analysis tools should account for different levels of grammaticality. Fuzzy grammaticality models may be a way to solve the problem that the so-called "noisy text" poses to parsing mechanisms used in Web language analysis–especially social networks language.

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE

Keywords: Grammaticality; parsing; web language; web data mining.

1. Introduction

This paper focuses on the problem that web language, especially social media text and the language used in computer mediated communication, poses to natural language processing, in particular to the traditional parsing techniques.

The development of information technologies, the rise of social media and the masive use of electronic devices in our daily communication have turned the web into a huge "data store" that attracts people that needs to obtain information about opinions and interests of the population. In this environment, fields such as Web Data Mining

* Corresponding author. Tel.: + 34 977 559 542; fax: + 977 558 386.
E-mail address: mariadolores.jimenez@urv.cat

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE
doi:10.1016/j.sbspro.2015.11.318

gain importance since the spectrum of applications is quite large and the economical and strategical interest is very high.

In order to extract information from the web we need to use natural language processing techniques since much of the content social media host is in the form of natural language. When we try to parse social media language with traditional parsing techniques we are in trouble. Parsing techniques have been designed to analyse grammatical constructions and to detect and reject ungrammatical input. Web language is not perfect, in general it is "bad language", it is full of grammatical violations or deviations. This non-cannonical input poses a big challenge to current parsing techniques that have to be modified in order to deal with this new type of language (Khan et al., 2013; McCloskys et al., 2012; Eisenstein 2013).

To solve the problem of parsing the web, different solutions have been proposed. *Normalization* of social media language and *domain adaptation* of current techniques are the two main computational approaches adopted to deal with "bad language" (Eisenstein, 2013). In this paper, we propose a different solution. We think that the problems that parsing techniques face when trying to analyse web language are a consequence of a traditional view of language in terms of discrete grammaticality, this is why we propose to solve the problem by changing this traditional conception through the introduction of new models that capture the idea of fuzziness in grammar by means of formal mechanisms.

2. Web Language

Before considering the problems that web language poses to natural language processing techniques, we have to examine the reasons that make social media language an increasingly important application domain for natural language processing. The answer to this question lies in the widespread and massive use that social networks have experienced recently. The 1350 million of active users who reached Facebook in 2014, the 500 million users of Twitter or the 260 million of LinkedIn are some examples of the rise of social media communication. These data demonstrate the importance and impact of social networks as a new interaction environment. As an essential tool for Internet users, social networks are a powerful media that attracts the attention of those who are interested in obtaining information about the population. For this reason, areas such as Web Data Mining are gaining importance nowdays.

Web Data Mining is defined as: "a broad class of software applications targeting at extracting information from Web sources" (Ferrara, 2014). Applications of this discipline are multiple, ranging from the analysis of documents of a company to bio-informatics, business and competitive intelligence and the crawling of online social networks. The importance of this research area depends on the fact that a large amount of information is continuously produced, shared and consumed online. Web data mining allows to efficiently collect this information with limited human effort. The obtained data is interesting from a sociological point of view –to learn about human behavior– and from a business viewpoint –enabling companies to acquire data about their competitors and to identify what are the topics catching the interest of its customers.

Much of the information hosted in the web appears in the form of text in natural language. Therefore, for the field of web data mining, natural language processing techniques are very important, especially parsing tools.

An important problem that web data mining has to deal with is the fact that social media language is produced in a natural and interactive way. Therefore, social media language shares with the spontaneous use of language grammatical deviations. This way of using language leads to the so-called "noisy text" (Baldwin et al., 2013).

In general, we can say that the features found in the web language are the same that define the spontaneous use of language. The fact that web language is written and not oral does not imply any difference. Hence, the grammar deviations rely on the spontaneous use of language regardless of the communication modality (written or oral). As stated by Hayes (1981):

> When people use natural language in natural conversation, they often do not respect grammatical niceties. Instead of speaking sequences of grammatically well-formed and complete sentences, people often leave out or repeat words or phrases, break off what they are saying and rephrase or replace it, speak in fragments, or use otherwise incorrect grammar.

The set of features identified by Hayes (1981) defines the social media language and turns it into “noisy text”. To assume this limitation in the web-language has led to those involved in its analysis to apply techniques of "shallow parsing" and to rule out other standard techniques of automatic processing of language, as stated by Baldwin (2013).

> Most research to date on social media text has used very shallow text processing (such as keyword-based time-series analysis), with natural language processing (NLP) tools such as part of-speech taggers and parsers tending to be disfavored because of the perceived intractability of applying them to social media text. However, there has been little analysis quantifying just how hard it is to apply NLP to social media text, or how intractable the data is for NLP tools.

3. Discrete grammaticality: A problem for web language

To analyse the data from the web by using standard parsing techniques is a very complex task because non-cannonical input still a challenge for natural language processing systems:

> Most natural language parsing algorithms are designed to analyze –clean— grammatical input. By the definition of their recognition process, these algorithms are designed to detect ungrammatical input at the earliest possible opportunity, and to reject any input that is found to be ungrammatical in even the slightest way. This property, which requires the parser to make a complete and absolute distinction between grammatical and ungrammatical input, makes such parsers fragile and of little value in many practical applications. Such parsers are thus unsuitable for parsing spontaneous speech, where completely grammatical input is the exception more than the rule. (Lavine, 1996).

This problem is related to the conception of language in terms of discrete grammaticality. Parsing algorithms are based on linguistic models that define grammaticality as a categorical notion: a sentence is grammatical or ungrammatical.

Although Chomsky (1975) acknowledged that "an adequate linguistic theory will have to recognize degrees of grammaticality”, linguists usually have prefered to consider grammaticality as a discrete notion. For example, Bouchard (1995) stated that “*fuzziness is not present in grammar in any way”.* Bever (1975) defended that:

> To give up the notion that a grammar defines a set of well-formed utterances is to give up a great deal. This is not to say that it is impossible in principle that grammars are squishy. Rather the possibility of studying precise properties of grammar and exact rules becomes much more difficult… Thus, if we can maintain the concept of discrete grammaticality, we will be in a better position to pursue an understanding of grammatical universals.

and Joos (1957) assumed that:

> Gradation or continuity in either form or meaning, has ever been found in any language of this planet. Nothing in language has degrees: everything is either this or that.

This sort of ideas has dominated theoretical linguistics and, somehow, has conditioned models in natural language processing. In fact, parsing techniques adopt a dichotomical conception of grammaticality rejecting any input that is ungrammatical in the slightest way.

4. Fuzzy grammaticality: A solution for web language?

Taking into account the difficulties that web language –especially, the language used in social networks– poses to standard parsers, Einsestein (2013) proposes two computational solutions:

- *Normalization techniques* that consist on modifying social media data to more closely resemble standard text. In this approach, in order to be able to parse web data, "bad" text is turned into "good" text by normalizing the language to better conform to the sort of input that traditional technology expects.
- *Domain adaptation* where rather than adapting text to fit existing tools, tools are adapted to the social media text.

We think that there is a third solution which does not consists on "normalizing" the text used in social networks and that does not require to develop specific tools just to analyze web-text. Our proposal claims that it is necessary to reformulate the theoretical grammatical models that are in the base of parsing algorithms.

Syntactic models should give up the idea of discreteness of grammar and adapt to the 'real' use of language where grammaticality is a matter of degrees. Therefore, we propose the definition of tools to formalize degrees of grammaticality in order to deal with the so-called "noisy text", which is very frequent not only in social media, but in every spontaneous use of language.

If natural language processing techniques have to deal with ungrammatical input, they should adopt a *new* concept of grammaticality already present in linguistics. Although, in general, theorical linguistics has considered grammaticality as a categorial notion, there are many linguists who have insisted on the necessity to defend the graduality of grammatical judgements in language (Pinker, 1999; Lakoff, 1973; Bolinger, 1961; Aarts, 2004; Aarts et al. 2004). The idea of fuzzy grammar is present in several linguistic models introduced from the nineies such as: *harmonic grammar*s (Smolensky and Legendre, 2006); *optimality theory* (Prince and Smolensky, 1993); *probability theory* (Manning, 2003); *property grammars* (Blache and Balfourier, 2001); and gradient *grammars* (Keller, 2000).

The above examples evidence the interest of dealing with natural language as a non-discrete object. We defend here the necessity of applying the ideas from fuzzy models of grammar to parsing algorithms techniques. In order to this, it is necessary to develop formal tools which capture the idea of fuzziness. Those formal tools should be based on speakers' acceptability/grammaticality judgments obtained from specific techniques in the area (Bader and Haussler 2010, Lau et al. 2014).

The design of parsers that understand grammaticality as a non-categorical concept would avoid both having to implement specific tools for parsing the language of social media and having to adapt the text of the web to the features of existing models. The same parsing algorithm could be used to process any kind of language (oral, written or web).

5. Final remarks

As stated by Eisenstein (2013), "the rise of social media has brought computational linguistics in ever-closer contact with bad language: text that defies our expectations about vocabulary, spelling, and syntax". Therefore, one of the immediate consequences of the popularity of social networks is to evidence the necessity to consider spontaneous language with their grammatical deviations and errors.

Social media language poses a big challenge to parsing algorithms. The problem of parsing the noisy language of social media opens a range of possibilities to natural language processing and leads linguists to reflect again, from a new perspective, on the problem of the notion of grammaticality.

This new scenario requires an interdisciplinary research where different areas –such as linguistics, formal language theory, psycholinguistics, computational linguistics, etc— are involved. The collaboration between those disciplines can provide a new theoretical framework on fuzzy grammaticality that combines formalization, linguistic theory and psycholinguistics evidence. The implementation of this model on parsing algorithms could become a good tool to improve, not only, the analysis of web language but the analysis of language in general.

References

Aarts, B. (2004). Conceptions of gradience in the history of linguistics. *Language Sciences,* 26, 343-389.

Aarts, B., Denison, D., Keizer, E., & Popova, G. (2004). *Fuzzy Grammar: A Reader*. Oxford: Oxford University Press.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46, 273–330.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text? How different Social Media Sources? In *Proceedings of 6th International Joint Conference on Natural Language Processing.* Nagoya.

Baldwin, T. (2012). Social media: Friend or foe of natural language processing? In *26th Pacific Asia Conference on Language, Information and Computation* (pp. 58–59).

Blache, P., & Balfourier, J.M. (2001). Property grammars: A flexible constraint-based approach to parsing. In *Proceedings of Seveth International Workshop on Parsing Technologies*. Beijing: Tsinghua University Press.

Bolinger, D.L. (1961). *Generality, Gradience and the All-Or-None*. The Hague: Mouton.

Bouchard, D. (1995). *The Semantics of Syntax: A Minimalist Approach to Grammar*. Chicago: University of Chicago Press.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. Nueva York: Plenum Press.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT 2013* (pp. 359-369). Atlanta.

Fanselow, G., Fery, C., Vogel, R., & Schelsewsky, M. (eds.) (2006). *Gradience in Grammar: Generative Perspectives*. Oxford: Oxford University Press.

Ferrara, E., De Meob, P., Fiumarac, G., & Baumgartnerd, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-based systems,* 70, 301–323.

Hayes, P. (1981). Flexible parsing. *American Journal of Computational Linguistics,* 7 (4), 232-242.

Joos, M. (1957). Description of Language Design. *Journal of the Acoustical Society of America,* 22, 701-708.

Keller, F. (2000). *Gradience in grammar: experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.

Khan, M., Dickinson, M., & Kübler, S. (2013). Does size matter? Text and grammar revision for parsing social media data. In *Proceedings of the Workshop on Language in Social Media* (pp. 1-10). Atlanta.

Lakoff, G. (1973). Fuzzy grammar and the perfomance/competence terminology game. In Corum, C.T., Smith-Stark, C., & Weiser, A. (Eds.), *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society* (pp. 271-291). Chicago: CLS.

Lau, J.H., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *The Annual Meeting of the Cognitive Science Society* (pp. 821-826).

Lavie, A. (1996). *GLR: A Robust Grammar-Focused Parser for Spontaneously Spoken Language*. PhD Thesis. Pittsburgh: Carnegie Mellon University.

Manning, C. (2003). Probabilistic approaches to syntax. In Bod, R., Hay, J., & Jannedy, S. (Eds.), *Probability Theory in Linguistics* (pp. 289-342). Cambridge: MIT Press.

McCloskys, D., Cheh, W., Recasens, M., Wangs, M., Sochers, R., & Manning, C.D. (2012). Stanford's system for parsing the English web. In *Notes of First Workshop on Syntactic Analysis of Non-Canonical Language.* Montreal.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. London: Widenfeld and Nicolson.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report, New Brunswick: Rutgers University.

Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge: MIT Press.

.