Experiments on Sentence Boundary Detection in User-Generated Web Content

Roque López and Thiago A. S. Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
{rlopez,taspardo}@icmc.usp.br

Abstract. Sentence Boundary Detection (SBD) is a very important prerequisite for proper sentence analysis in different Natural Language Processing tasks. During the last years, many SBD methods have been used in the transcriptions produced by Automatic Speech Recognition systems and in well-structured texts (e.g. news, scientific texts). However, there are few researches about SBD in informal user-generated content such as web reviews, comments, and posts, which are not necessarily well written and structured. In this paper, we adapt and extend a well-known SBD method to the domain of the opinionated texts in the web. Particularly, we evaluate our proposal in a set of online product reviews and compare it with other traditional SBD methods. The experimental results show that we outperform these other methods.

Keywords: Sentence Boundary Detection, Noisy Text Processing, User Generated Content

1 Introduction

In the last decade, many websites have appeared where users may freely generate content and with few restrictions. Websites such as forums, wikis and product review sites have become big repositories of information about different topics. Unfortunately, in these websites, the vast majority of this information is usually written in an informal and, sometimes, ill-formed way, not following orthography and grammar rules. For instance, in product reviews, it is very common to find a lot of noise, such as spelling mistakes, non-standard abbreviations and missing or inadequate sentence boundary marks [9].

Sentence Boundary Detection (SBD) is the focus of this paper. This task consists in identifying the sentences within a text [26]. In Automatic Speech Recognition (ASR), it is very popular due to the necessity of finding sentential segments in the stream of words (the transcripts) that are automatically recognized. In text processing, it is essential to produce the input – the sentences – to other tools (as POS tagger and parser) and applications (as information extraction and summarization).

In the majority of the languages, the period (".") is usually employed as sentence boundary marker, while it may also be used in abbreviations, acronyms,

ordinal numbers, e-mails and URLs. The variety of applications of the period mark represents a challenge in the SBD task. In online user-generated content, this challenge is even greater because this marker is usually omitted or not properly used. Figure 1 shows an example of a (real) product review written in Brazilian Portuguese and translated into English. As we may see, users generally do not use the period mark to delimit sentences as well as do not respect the use of other punctuation marks (as commas and semicolons) or capital letters, making more challenging the SBD process and, consequently, the other tasks that depend on it.

```
Prós: O telefone deve ser ótimo

Contras: Cuidado com a Empresa_X...tem preço bom
mas péssima entrega (Empresa_Y é palhaçada)

Opinião: Não recomendo a ninguém comprar na
Empresa_X;nop e-commerce eles são piores que o
Empresa_Z


[Possible translation]

Pros: The phone must be great

Cons: Beware the Company_X...it has good price
but bad delivery (Company_Y is a joke)

Opinion: I do not recommend anyone to buy in
Company_X;in the e-commerce they are worse than
the Company_Z
```

Fig. 1. Example of online product review[1]

To demonstrate the relevance of tackling such issues, [9] recently presented an analysis of different kinds of noise in online product reviews written in Brazilian Portuguese. In that study, the manual correction of punctuation marks led to an improvement of 4.34% in the precision of the POS tagger.

In this paper, we explore SBD methods in user-generated web content. We start by adapting and extending the supervised machine learning method proposed in [25]. This is one of the most classical methods and, unlike other ones, does not use prosodic information (e.g., rhythm, stress, or intonation) – as it usually happens in the ASR context – and thus it is suitable for written texts. We also evaluate two other SBD systems, MxTerminator [22] and Punkt [11], which are considered state of the art systems. In particular, for training the machine learning method, we use well-written news texts, expecting that patterns for good usage of period mark may be learned and used for SBD in user-generated content.

[1] Company names were omitted in this figure due to ethical concerns.

We opted to run our experiments on texts of the same corpus used in [9], which is composed of product reviews written in Brazilian Portuguese, retrieved from a product evaluation webpage. We agree with [9] that such texts are good representatives of the writing phenomena that occurs in user-generated web content. Finally, as our database is in Portuguese, we use some corpora of news texts in this language for training the machine learning solution, adopting, in the end, the publicly available CSTNews corpus [7].

We show that our results outperformed the other state of the art methods and, interestingly, that a large training corpus is not necessary for achieving good results. The remaining of the paper is organized as follows: in Section 2, we introduce the main related work; in Section 3, we describe the proposed method to identify sentence boundaries; the experiments and results are presented in Section 4; finally, in Section 5, we conclude this paper.

2 Related Work

There are many approaches used to detect sentence boundaries in different languages. According to [24], SBD systems are grouped in two classes: methods that use fixed rules and methods that use machine learning techniques. These methods have been well studied in the ASR area and are widely applied in news texts [5][11][21][25]. In this section, we comment some of these methods.

For Brazilian Portuguese language, [24] is one of the first works in SBD, with very interesting results. The authors compare the performance of two systems that use machine learning methods (MxTerminator [22] and Satz [16]) and one system based on fixed rules (RE SYSTEM [23]). These systems were evaluated in a corpus of news texts in two scenarios: (i) when the domain of the texts is known in advance, the results of these systems were similar, and (ii) when the domain is unknown, the best results were obtained by the machine learning methods. The main reason for these results is that rules are dependent on the domain.

MxTerminator [22], tested in the work above mentioned, and Punkt [11] are language-independent SBD systems and have been used in many languages, including Portuguese. MxTerminator uses a statistical approach based on Maximum Entropy to identify the sentences of a document. From a corpus with the sentences already identified, this method learns the contextual information where sentence boundaries occur. For this, MxTerminator uses some features, such as the preceding token, the following token, and capitalization information. For Brazilian Portuguese, MxTerminator showed a robust performance (96.46 of F-measure) in the Lacio-Web Corpus [3] using 10-fold cross-validation. Punkt is an unsupervised SBD system based on the assumption that, once abbreviations have been identified, it is more feasible to identify sentence boundaries. For this, Punkt uses properties of abbreviations to identify them and considers that all periods not attached to an abbreviation are sentence boundaries. Additionally, Punkt uses some heuristics (e.g., the presence of digits followed by a period mark) to identify name initials and ordinal numbers. For Brazilian Portuguese,

Punkt outperformed the results of MxTerminator with 97.22 of F-measure in the same corpus (Lacio-Web).

[17] presents SENTER, a rule-based system to sentence segmentation of well written texts. This system is very simple and uses some general heuristics to detect sentence boundaries (such as the presence of newline characters and the different possibilities where the period mark is not a sentence boundary symbol). In that work, the authors do not present an evaluation about the performance of SENTER.

In the ASR area, [5] made experiments concerning punctuation and capitalization recovery for spoken texts about news in European Portuguese. In order to recover the period mark, the authors use maximum entropy models with some features like n-grams, POS tags and prosodic information. In the experiments, the authors show that lexical features had less impact than prosodic features, but the combination of all features produced better results.

For informal user-generated content, there are few researches on SBD. [21] evaluated several SBD systems in news texts and user-generated content written in English. As expected, the lowest results were obtained in informal texts, because, according to the authors, in these texts there is a decline in linguistic formality. For Brazilian Portuguese, as far as we know, there is still no SBD works for informal texts. For others languages, like Arabic and Chinese, there are some efforts. [1] uses common words as sentence delimiting symbols in Arabic texts, and [27] presents a maximum entropy model-based approach to predict and correct punctuation marks to segment sentences written in Chinese.

In this paper, we test MxTerminator and Punkt in the intended scenario and compare their results with the main method that we explore in this paper, which we introduce in what follows.

3 Our Approach

The proposed approach in this study is an adaptation of the supervised machine learning method proposed in [25]. In that work, the authors introduced the problem of SBD on the text produced by ASR systems and used written texts to evaluate their proposal. The authors used the Timbl memory-based learning algorithm [8] with a set of twelve features derived from the analysis of the preceding and following words in relation to the point where the punctuation should be included. To train and test their method, they used news of the Wall Street Journal.

Before detailing the features and our approach, it is important to clarify how to model the task as a machine learning solution. According to [12], the SBD problem may be represented as a classification task in the following way: for each word in the text, we determine whether it is or not a sentence boundary, i.e., each word (the learning instance) might be classified as belonging to either the *boundary* class or the *no_boundary* class. Words of the *boundary* class are those that should be followed by a period mark. This is the general learning schema used in [25] and is adopted in this work.

In our proposal, in addition to consider the twelve features used in [25], we experiment an extended version with two more features: (i) a flag indicating whether the following token is a newline mark and (ii) a flag indicating whether the following token is a period mark. In Table 1, we show the fourteen features used in this paper: the first twelve features are those used in [25] and the last two features are the ones proposed above. While some of these features are computed in traditional ways, some deserve explanations.

Table 1. Features used in the proposed approach

Id	Feature
F1	The preceding word
F2	Probability that the preceding word ends a sentence
F3	Part of speech tag assigned to the preceding word
F4	Probability that the above part of speech tag (feature F3) is assigned to the last word in a sentence
F5	Flag indicating whether the preceding word is a stopword
F6	Flag indicating whether the preceding word is capitalized
F7	The following word
F8	Probability that the following word begins a sentence
F9	Part of speech tag assigned to the following word
F10	Probability that the above part of speech tag (feature F9) is assigned to the first word in a sentence
F11	Flag indicating whether the following word is a stopword
F12	Flag indicating whether the following word is capitalized
F13	Flag indicating whether the following token is a period mark
F14	Flag indicating whether the following token is a newline mark

We propose the newline mark as a feature because, in product reviews, users usually use this symbol as a sentence boundary. It is very common in online informal texts. In the case of the period mark, we consider this feature because, although they are rarely used, when users use this symbol, it is very likely that it is a sentence delimiter. For this feature, using regular expressions, we previously filter out occurrences of period marks that are decimal points or parts of e-mails and URLs.

For capitalized words (feature F6 and F12), we perform a simple analysis. We verify if the first letter is the only capitalized letter, because, in product reviews, users do not respect the correct use of capitalized words. Cases like *PRODUTO RUIM* (BAD PRODUCT, in English) or *BoM SeRvIÇo* (GoOd SeRvIcE, in English) are very common. These examples, with mixed letter cases, we consider as lowercase words. We believe that when users employ these types of words they want to highlight an expression and not to start a new sentence.

For features F3 and F9, [25] used the POS tags manually annotated in the news of the Wall Street Journal. In our case, the product reviews do not present previously annotated POS tags. For this reason, we followed a probabilistic approach. This approach uses as data source the Mac-Morpho corpus [2], in which

each word shows the corresponding correct POS tag. To tag a word in product reviews, our approach searches the most likely tag for that word, i.e., the tag that is the most used one in the above corpus. For words not present in MacMorpho, we used the first listed tag in the DELAF dictionary [15], in which each word is associated to all its possible tags. In the case a word is not present in both sources, we consider it as a noun. As an alternative, a traditional POS tagger might be used, but, in product reviews, there are many noises that affect the performance of POS taggers. This motivated us to use the probabilistic approach.

Once we have the features, we used Naïve Bayes as the machine learning method, specifically the version implement in scikit-learn library [18]. We also conducted some experiments with others machine learning methods (SVM, k-Nearest Neighbors and Stochastic Gradient Descent), but Naïve Bayes got the best results. For this reason, we only report its results. As said before, we train our method with well-written news texts, expecting that we may learn patterns of good usage of period marks to detect (in the test phase) where user-generated texts need segmentation. We describe the corpora we tested in the next section. It is also important to say that it was not possible to use user-generated texts for training our method because there is no such manually annotated data available, to the best of our knowledge.

As input, our method receives a product review in plain text format. After that, we eliminate all punctuations marks and, for each word in the text, we extract the features showed in Table 1. With these features, our proposal determines whether the word evaluated is at a sentence boundary position (and should have a period mark inserted after it) or not. Finally, an output is generated with the detected sentence boundaries. In Figure 2, we show an example of input and output of our method. The input and output are in the top and bottom of the figure, respectively.

```
Prós: O telefone deve ser ótimo

Contras: Cuidado com a Empresa_X...tem preço bom
mas péssima entrega (Empresa_Y é palhaçada)

Opinião: Não recomendo a ninguém comprar na
Empresa_X;nop e-commerce eles são piores que o
Empresa_Z
```

```
Prós: O telefone deve ser ótimo.

Contras: Cuidado com a Empresa_X... tem preço
bom mas péssima entrega (Empresa_Y é palhaçada).

Opinião: Não recomendo a ninguém comprar na
Empresa_X; nop e-commerce eles são piores que o
Empresa_Z.
```

Fig. 2. Examples of input and output data in our proposed approach

4 Experiments

4.1 Datasets

To train the proposed method, we used the CSTNews corpus [7], a collection of news texts written in Brazilian Portuguese. As CSTNews is a small corpus, we initially did experiments with much larger corpora, using the Corpus NILC [19] and PLN-Br GOLD corpus [6] in the training phase, but, surprisingly, the results were not better. More than this, in Corpus NILC there were some sentences (tittles of the news) without period marks, and this affected the learning process. In the case of PLN-Br GOLD corpus, the results were similar but it took a long time to process all the documents.

As the results were not better with the larger corpora, we believe that our proposal have a good learning process with few data. For these reasons, we only used the CSTNews corpus in the training phase and the results we report here are based on this corpus.

The CSTNews is a corpus composed of 140 news texts grouped in 50 clusters. Each cluster contains from 2 to 3 news texts on the same topic compiled from some of the main online newspapers in Brazil. These texts are news about sports, politics, science and others. In total, this corpus has 2067 sentences. Additionally, CSTNews has other types of manual annotations, like CST (Cross-document Structure Theory) [20], RST (Rhetorical Structure Theory) [14], multi-document summaries and their alignment with the corresponding source texts, among other annotation layers. In this study, we only use the full texts of this corpus.

To test our methods, we used the corpus of product reviews described in [9], which were collected from Buscapé[2], a website where users comment about different products (e.g., smartphones, digital cameras, notebooks, etc.). These comments are written in a free format within a template with three sections: Pros, Cons, and Opinion.

To conduct the experiments, we used a sample of 35 product reviews annotated by a computational linguist. The annotation consisted in deletions, insertions or substitutions of punctuations marks to correct the texts. This data was all the data that we had available for testing the methods.

4.2 Results

In the experiments, we evaluated our proposal, the original method proposed by [25], and two state-of-the-art SBD systems: MxTerminator [22] and Punkt [11], which have the highest results reported in the literature [26]. For the experiments, we use the implementations of OpenNLP [4] and NLTK [13] libraries for MxTerminator and Punkt, respectively. We also tried to use the sentence separator proposed by [26], but the online version was not working for Portuguese texts.

Table 2 shows the results of our experiments with Precision, Recall and F-measure metrics. Precision is computed as *tp / (tp + fp)*, where *tp* is the number

[2] http://www.buscape.com.br/

of true positives and *fp* the number of false positives. Recall is the ratio *tp / (tp + fn)*, where *tp* is the number of true positives and *fn* the number of false negatives. F-measure is the harmonic mean of precision and recall, being a unique indicator of the quality of the method. The overall results shown in Table 2 are the averages over the *boundary* and *no_boundary* classes. One may see that our proposal obtained the best results.

Table 2. Overall results

Method	**Precision**	**Recall**	**F-Measure**
MxTerminator [22]	0.939	0.847	0.886
Punkt [11]	0.943	0.843	0.885
Original Approach [25]	0.801	0.834	0.817
Proposed Approach	**0.953**	**0.895**	**0.921**

With the use of news texts in the training process, the results were good, showing that good patterns could be learned, as we had hypothesized before. We may also see that the results obtained by our proposal are better than the original method proposed by [25] in Precision, Recall and F-measure, reflecting that our two additional features (F13 and F14) helped improving the performance of the method.

It is important to highlight that, using the Student's t-test with 95% of confidence, the differences between the F-measures obtained by our proposal and the other methods are statistically significant. In relation to the general accuracy, our proposal also got the best results, with 97.60%, while the original method achieved 93.70%, and MxTerminator and Punkt 96.70%.

We used the Relief algorithm [10] to evaluate the importance of each feature of our proposal. Of the fourteen features used, the probability that the POS tag is assigned to the first word in a sentence (F10) and the POS tag assigned to the following word (F9) do not contribute to the final performance. In other words, removing these features does not affect the results. However, if we remove any other feature, the performance decreases. The three best features were the presence of the newline mark (F14), the probability that the following word begins a sentence (F8) and the presence of the period mark (F13).

In order to analyze the performance of our proposal in more details, we show, in Table 3, the results obtained by the four SBD systems for the words that belong to the *boundary* class (the *yes* class, therefore). Clearly, the proposed approach in this paper got the best results.

We attribute these best results to the special analysis that we made of the product reviews characteristics, such as usage of capitalization, use of newline marks and POS tags in informal texts. On the other hand, a very common error made by our method in identifying boundaries occurred when words are unknown. These new words are not present in the training corpus and our proposal cannot learn patterns when these words are followed by the period mark. These new words may simply be unseen words in the training data as well as

Table 3. Results for the *boundary* class

Method	Precision	Recall	F-Measure
MxTerminator [22]	0.907	0.701	0.791
Punkt [11]	0.915	0.693	0.789
Original Approach [25]	0.632	0.708	0.668
Proposed Approach	**0.925**	**0.795**	**0.855**

spelling mistakes, foreign words or slangs, which are typical elements in product reviews.

We believe that, if we use an annotated corpus of reviews in the training phase, these types of errors would not frequently occur because the machine learning method would identify and learn, with more coverage, the features of this domain. In this context, Recall measure, that was low (see Table 3), would improve.

It was also evaluated the performance of the SBD systems for the *no_boundary* class (the *no* class). The obtained results are presented in Table 4. In comparison with Table 3, the performances are much better and there is little difference among the four methods. It is because the majority of words in texts are not sentence boundaries, and, thus, there are more instances of the *no_boundary* class in the training phase. We believe that this unbalance in the data influenced the results for this class.

Table 4. Results for the *no_boundary* class

Method	Precision	Recall	F-Measure
MxTerminator [22]	0.971	0.993	0.982
Punkt [11]	0.971	**0.994**	0.982
Original Approach [25]	0.971	0.960	0.965
Proposed Approach	**0.980**	**0.994**	**0.987**

In relation to other romance languages, such as Spanish or French, we believe that it is possible to use the fourteen features of our proposal and get satisfactory results, because these languages share some common linguistics characteristics like the basic *subject–verb–object* order. In addition, we believe that internet users of these languages have similar behavior when they generate web content (e.g., use of newline marker). However, for other languages, such as Chinese or Japanese, it is complicated to use our approach because their linguistic characteristics are different and some of our machine learning features are not present in these languages, such as the capitalization rule (features F6 and F12).

5 Conclusion and Future Work

In this work, we analyzed the SBD problem in user-generated web content. As it may be seen, we adapted and extended a classical approach to the problem and

outperformed other state of the art systems. This research has been motivated, mainly, by the importance of the SBD systems in the preprocessing of web texts for posterior processing by other NLP tools.

As a future work, we plan to study the use of the above methods for detecting other punctuation marks, as comma and semicolon, which must be bigger challenges to deal with, since their usage is more flexible in several different situations.

Acknowledgments. Part of the results presented in this paper were obtained through research on a project titled "Semantic Processing of Texts in Brazilian Portuguese", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

1. Al-Subaihin, A., Al-Khalifa, H., Al-Salman, A.: Sentence Boundary Detection in Colloquial Arabic Text: A Preliminary Result. In: Proceedings of the International Conference on Asian Language Processing. pp. 30–32 (2011)
2. Aluísio, S., Pelizzoni, J., Marchi, A., de Oliveira, L., Manenti, R., Marquiafável, V.: An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: Computational Processing of the Portuguese Language. Lecture Notes in Computer Science, vol. 2721, pp. 110–117 (2003)
3. Aluísio, S., Pinheiro, G., Finger, M., Nunes, M.G., Tagnin, S.: The LacioWeb Project: Overview and Issues in Brazilian Portuguese Corpora Creation. In: Proceedings of Corpus Linguistics. pp. 14–21 (2003)
4. Baldridge, J.: The OpenNLP Project. `http://opennlp.apache.org/index.html` (2005), [Accessed 15 January 2015]
5. Batista, F., Caseiro, D., Mamede, N., Trancoso, I.: Recovering Capitalization and Punctuation Marks for Automatic Speech Recognition: Case Study for Portuguese Broadcast News. Speech Communication 50(10), pp. 847–862 (2008)
6. Bruckschen, M., Muniz, F., Souza, J., Fuchs, J., Infante, K., Muniz, M., Gonçalves, P., Vieira, R., Aluísio, S.: Anotação Linguística em XML do Corpus PLN-BR. Série de Relatórios do NILC, NILC-TR-09-08 (2008)
7. Cardoso, P.C., Maziero, E.G., Jorge, M., Seno, E.M., Di Felippo, A., Rino, L.H., Nunes, M.d.G.V., Pardo, T.A.: CSTNews-A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: Proceedings of the 3rd RST Brazilian Meeting. pp. 88–105 (2011)
8. Daelemans, W., Jakub, Z., Van Der Sloot, K., Van Den Bosch, A.: TiMBL: Tilburg Memory Based Learner-Version 2.0 - Reference Guide (1999)
9. Duran, M., Avanço, L., Aluísio, S., Pardo, T., Nunes, M.d.G.: Some Issues on the Normalization of a Corpus of Products Reviews in Portuguese. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9). pp. 22–28 (2014)
10. Kira, K., Rendell, L.A.: The Feature Selection Problem: Traditional Methods and a New Algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence. pp. 129–134 (1992)
11. Kiss, T., Strunk, J.: Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32(4), pp. 485–525 (2006)

12. Liu, Y., Chawla, N.V., Harper, M.P., Shriberg, E., Stolcke, A.: A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech. Computer Speech & Language 20(4), pp. 468–494 (2006)
13. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. pp. 63–70 (2002)
14. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: A Theory of Text Organization. University of Southern California, Information Sciences Institute (1987)
15. Muniz, M.C., Nunes, M.D.G.V., Laporte, E.: UNITEX-PB, a set of Flexible Language Resources for Brazilian Portuguese. In: Workshop on Technology on Information and Human Language. pp. 2059–2068 (2005)
16. Palmer, D.D., Hearst, M.A.: Adaptive Multilingual Sentence Boundary Disambiguation. Computational Linguistics 23(2), pp. 241–267 (1997)
17. Pardo, T.A.S.: SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. Série de Relatórios do NILC, NILC-TR-06-01 (2006)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research 12, pp. 2825–2830 (2011)
19. Pinheiro, G.M., Aluísio, S.M.: Corpus Nilc: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web. Série de Relatórios do NILC, NILC-TR-06-09 (2003)
20. Radev, D.R.: A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure. In: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue. pp. 74–83 (2000)
21. Read, J., Dridan, R., Oepen, S., Solberg, J.L.: Sentence Boundary Detection: A Long Solved Problem? In: Proceedings of 24th International Conference on Computational Linguistics. pp. 985–994 (2012)
22. Reynar, J.C., Ratnaparkhi, A.: A Maximum Entropy Approach to Identifying Sentence Boundaries. In: Proceedings of the 5th Conference on Applied Natural Language Processing. pp. 16–19 (1997)
23. Silla, C., Kaestner, C.: Automatic Sentence Detection Using Regular Expressions (in Portuguese). In: Proceedings of the 3rd Brazilian Computer Science Congress. pp. 548–560 (2003)
24. Silla, C., Kaestner, C.: An Analysis of Sentence Boundary Detection Systems for English and Portuguese Documents. In: Computational Linguistics and Intelligent Text Processing. pp. 135–141 (2004)
25. Stevenson, M., Gaizauskas, R.: Experiments on Sentence Boundary Detection. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. pp. 84–89 (2000)
26. Wong, D.F., Chao, L.S., Zeng, X.: iSentenizer-μ: Multilingual Sentence Boundary Detection Model. The Scientific World Journal 2014 (2014)
27. Zhao, Y., Fu, G.: A MEMs-based Labeling Approach to Punctuation Correction in Chinese Opinionated Text. In: Proceedings of the 2013 International Conference on Intelligence Artificial. pp. 329–336 (2013)