# Disambiguation of Period Characters in Clinical Narratives

**Markus Kreuzthaler** and **Stefan Schulz**
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
<markus.kreuzthaler,stefan.schulz>@medunigraz.at

## Abstract

The period character's meaning is highly ambiguous due to the frequency of abbreviations that require to be followed by a period. We have developed a hybrid method for period character disambiguation and the identification of abbreviations, combining rules that explore regularities in the right context of the period with lexicon-based, statistical methods which scrutinize the preceding token. The texts under scrutiny are clinical discharge summaries. Both abbreviation detection and sentence delimitation showed an accuracy of about 93%. An error analysis demonstrated potential for further improvements.

## 1 Introduction

The full stop, or period character, is ambiguous. As well as its use as a sentence delimiter, it is often collocated with abbreviations ("Prof."), occurs in numeric expressions ("13.2 mg"), including dates, and appears in a series of special names such as Web addresses. Minor variations exist between languages and dialects (for example the use of the period as decimal delimiter), and rule variations exist that guide its collocation with abbreviations. The character-wise analysis of text can produce a clear distinction between (i) period characters that are enclosed between two alphanumeric characters, and (ii) period characters that are adjacent to at least one non-alphabetic character. Whereas in the former case the period character can be considered an internal part of a token, the latter allows for two interpretations:

1. Period characters that are mandatorily collocated with abbreviations; and
2. Period characters as sentence delimiters.

We focus on text produced by physicians at the point of care, either directly or via dictation. The sublanguage of clinical narratives is characterized, among other peculiarities such as misspellings, punctuation errors, and incomplete sentences, by the abundance of acronyms and abbreviations (Meystre et al., 2008). It is for this reason that we focus here on the use of the period character to distinguish between sentence limits and abbreviations.

A snippet from a medical text illustrates some typical phenomena:

```
  3. St.p. TE eines exulz.
sek.knot.SSM (C43.5) li Lab.
majus. Level IV, 2,42 mm
Tumordurchm.
```

In "3." the period marks an ordinal number; "St.p." is the abbreviation of "Status post" (state after); "TE" is an acronym derived from "Totale Exzision". "Exulz." and "Tumordurchm." are ad-hoc abbreviations for "exulzerierendes" and "Tumordurchmesser" (tumour diameter), respectively. "sek.knot.SSM" is an ill-formed agglutination of two abbreviations and one acronym. In correctly formatted text, they would be separated by spaces ("sek. knot. SSM"). The abbreviation "sek." (secondary) is written in a common lexicalized form, whereas "knot." is, once again, an ad-hoc creation. "SSM" is an acronym for "Superfiziell Spreitendes Melanom". "C43.5" is a code from the International Classification of Diseases[1]. "Lab." means "Labium", a common anatomical abbreviation. "IV" is not an acronym, but a Roman number. "2,42" is a decimal number, demonstrating that the comma rather than the period is used as a decimal separator in German texts. Finally, the abbreviation "Tumordurchm." exemplifies that

[1] http://www.who.int/classifications/icd/en/



*Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014*, pages 96–100,
Gothenburg, Sweden, April 26-30 2014. 

the period can play a double role, *viz.* to mark an abbreviation and to conclude a sentence.

In this paper we will describe and evaluate a methodology that is able to identify and distinguish the following: (i) periods that act as sentence delimiters after ordinary words (such as the period after "majus") marked as **NSD** (normal sentence delimiter); (ii) periods as abbreviation markers in the middle of a sentence, marked as **MAM** (mid-sentence abbreviation marker), and (iii) periods that are both abbreviation markers and sentence delimiters, marked as **EAM** (end-sentence abbreviation marker). From this ternary distinction, two binary tasks can be derived, *viz.* the detection of abbreviations (MAM and EAM), and the detection of sentence endings (NSD and EAM).

## 2 Materials and Methods

### 2.1 Data

We used 1,696 discharge summaries extracted and anonymized from a clinical information system. They had an average word count of 302, with a mean of 55 period characters per document. The texts were divided into a learning set (1.526 documents) and an evaluation set (170 documents). Two word lists were created in advance: (i) a medical domain dictionary (MDDict) with a high coverage of domain-specific terms, excluding abbreviations, and (ii) a closed-class dictionary (CCDict) containing common, domain-independent word forms.

For **MDDict**, words were harvested from three sources: a free dictionary of contemporary German[2], a word list created out of raw text extracted from a medical dictionary on CD-ROM (Pschyrembel, 1997), and medical texts and forum postings from a patient-centered website[3]. The final list comprised approximately 1.45 million types, which were subsequently indexed with Lucene[4]. This dictionary was modified during a second step by two Web resources containing German abbreviations[5,6]. We accumulated about 5,800 acronym and abbreviation tokens, which were then removed from the Lucene-indexed dictionary, in order to transform MDDict into a resource mostly devoid of abbreviations.

[2]http://sourceforge.net/projects/germandict/
[3]http://www.netdoktor.at/
[4]https://lucene.apache.org/core/
[5]http://de.wikipedia.org/wiki/Medizinische_Abkürzungen
[6]http://de.wiktionary.org/wiki/Kategorie:Abkürzung

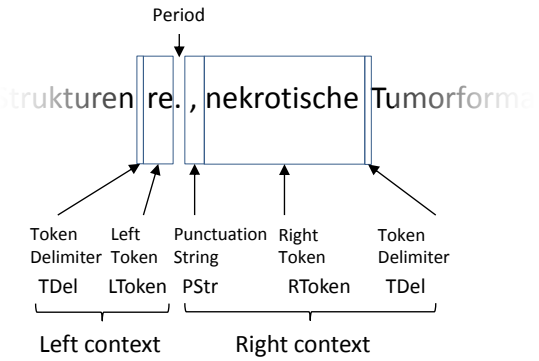


Figure 1: Period pattern and zoning of left and right context.

For **CCDict** we harvested closed-class words from a German web resource[7], i.e. prepositions, determiners, conjunctions, and pronouns, together with auxiliary and modal verbs. The purpose of this was to arrive at a comprehensive list of word forms that can only be capitalized at the beginning of a sentence.

Figure 1 shows the pattern used to identify periods of interest for this study. The right and the left context were zoned as followed: The string to the left of the period until the preceding token delimiter is the "Left Token" (**LToken**). The sequence of spaces, line breaks, or punctuation marks to the right of the period ("Punctuation String") is identified as **PStr**. The following token, spanning from the first alphanumeric character to the character left to the next delimiter, is named **RToken**.

### 2.2 Context evaluation

The right context is evaluated first (Algorithm 1). It is based on the following assumptions: (i) Whenever a period terminates a sentence, the first character in the following token is capitalized. For a subset of words this can be ascertained by looking up the closed word class dictionary CCDict (the restriction to "closed classes" is due to the fact that German nouns are mandatorily capitalized, including nominalized adjectives and verbs); (ii) A sentence can never be split by a line break, therefore a period that precedes the break necessarily marks the end of the previous sentence; (iii) Most punctuation signs that follow a period strongly indicate that the period character here plays the role of an abbreviation marker and does not coincide with an end-of-sentence marker. Only in the case where a decision could not be achieved using the

[7]http://www.deutschegrammatik20.de/

```
if RToken begins with lower case character
then
    → MAM;
else
    if decapitalized RToken matches closed
    class token then
        → EAM or NSD;
    else
        if If PStr contains punctuation
        character then
            → MAM;
        else
            if If PStr contains a line break
            then
                → NSD or EAM;
            else
                → NSD or MAM or EAM;
            end
        end
    end
end
```

**Algorithm 1:** Rule-based decision algorithm for the right context of a period.

algorithm is the left context investigated.

The evaluation of the left context extends the approach from Kiss and Strunk (2002), who used the *log likelihood ratio* (Dunning, 1993) for abbreviation detection:

$$log\lambda = -2log(L(H_0)/L(H_A))$$

$H_0$ is the hypothesis that the occurrence of a period is independent of the preceding word, $H_A$ the hypothesis that it is not independent.

We use four scaling functions $S_1$ – $S_4$. The period character is symbolized by $\bullet$; $C(word, \bullet)$ and $C(word, \neg\bullet)$ describe the co-occurrence frequency counts. The primary $log\lambda$ is modified by sequential composition. Following Kiss and Strunk (2002), $S_1$ enhances the initial $log\lambda$ if $C(word, \bullet)$ is greater than $C(word, \neg\bullet)$. $S_2$ varies from $-1$ to 1 depending on $C(word, \bullet)$ and $C(word, \neg\bullet)$. $S_3$ leads to a reduction of $log\lambda$ depending on the length of the preceding word. We introduced a fourth scaling function $S_4$, which reflects the fact that most abbreviations are proper substrings of the shortened original word (e.g. "exulz." = "exulzerierend"), with $N$ being the sum of all found substring matches in the form $subword_i*$ for every $subword_i$ in $subword_1 \bullet subword_2 \bullet \ldots subword_n\bullet$ in a Lucene search result.

$$S_4(log\lambda) : log\lambda + N(word, \bullet)$$

This also includes those abbreviations which have an internal period, such as "St.p". The reason why the last scaling function contains an addition, is to accommodate for cases where $C(word, \bullet) < C(word, \neg\bullet)$ even when $word$ is an abbreviation. These cases, for which the weighted $log\lambda$ is negative, could then nevertheless be pushed to the positive side in the result of a strong $S_4$.

For the final decision in favor of an abbreviation, we required that the following two conditions hold: (i) $(S_1 \circ S_2 \circ S_3 \circ S_4)(log\lambda) > 0$; (ii) the length of the abbreviation candidate was within the 95% confidence interval, given the statistical distribution of all abbreviation candidates that exhibited a significant collocation ($p < 0.01$), $C(word, \bullet) > C(word, \neg\bullet)$, and MDDict not containing $word$.

## 3 Results

For the evaluation methodology, a gold standard was created by a random selection of 500 text frames, centered around a period with its left and right context (each 60 characters) from the evaluation set. The two authors rated each period in the center of the snippet as being an NSD, a MAM or an EAM. A subset of 100 was rated by both authors in order to compute the inter-rater agreement. We obtained a Cohen's kappa (Di Eugenio and Glass, 2004, Hripcsak and Heitjan, 2002) of 0.98, when rating both abbreviation vs. non-abbreviation, and sentence delimiter vs. non sentence delimiter, respectively. Accuracy, true and false negative rates (Manning et al., 2008), are computed for the two processing steps in isolation. This required making some default assumptions for the cases in which the result was ambiguous. The assumptions are based on frequency distributions of the three values in the learning set. The left context processing detects abbreviations, but is unable to distinguish between EAM and MAM. As the frequency of MAM is much higher, this value is set wherever NSD is discarded. In the processing of the right context, the algorithm may fail to disambiguate between NSD vs. EAM, or even terminate with any decision (NSD vs. EAM vs. MAM), cf. Algorithm 1. In the latter case MAM is set, as this was determined to be the most frequent phenomenon in the learning data (0.53). In

the former case, NSD is given preference over EAM, which has a low frequency in the learning set (0.03). Table 1 shows accuracy and false positive / negative rates obtained by left, right and combined context evaluations.

| | Accuracy | Fpos | Fneg |
|---|---|---|---|
| *Abbreviation detection* | | | |
| Left | 0.914 | 0.035 | 0.136 |
| Right | 0.880 | 0.162 | 0.051 |
| L & R | 0.928 | 0.060 | 0.082 |
| *Sentence delimitation* | | | |
| Left | 0.902 | 0.107 | 0.077 |
| Right | 0.884 | 0.014 | 0.211 |
| L & R | 0.934 | 0.062 | 0.065 |

Table 1: Abbreviation detection and sentence delimitation results.

It is remarkable that the combination of both algorithms only produces a moderate gain in accuracy. For the minimization of certain false negatives and false positives, it can be advantageous to consider the right or left context separately. For instance, the right context algorithm alone is better at minimizing false positive sentence recognitions, whereas the left context algorithm is better suited at minimizing cases of false positive abbreviation detections. Apart from known issues such as the above mentioned parsing problems, for which the reader needs to be familiar with the domain and the style of the documents, the analysis of misclassifications revealed several weaknesses: sensitivity to spelling and punctuation errors (especially missing spaces after periods) and abbreviations that can also be read as a normal word (e.g. “Mal.” for “Malignität” or “Mal” (time)), and abbreviations that are still present in MDDict.

## 4 Related Work

The detection of short forms (abbreviations, acronyms) is important due to their frequency in medical texts (Meystre et al., 2008). Several authors studied their detection, normalization, and context-dependent mapping to long forms (Xu et al., 2012). CLEF 2013 (Suominen et al., 2013) started a task for acronym/abbreviation normalization, using the UMLS[8] as target terminology. An F-Measure of 0.89 was reported by Patrick et al. (2013). Four different methods for abbreviation detection were tested by Xu et al. (2007). The fourth method (a decision tree classifier), which additionally used features from knowledge resources, performed best with a precision of 91.4% and a recall of 80.3%. Therefore Wu et al. (2011) compared machine learning methods for abbreviation detection. Word formation, vowel combinations, related content from knowledge bases, word frequency in the overall corpus, and local context were used as features. The random forest classifier performed best with an F-Measure of 94.8%. A combination of classifiers lead to the highest F-Measure of 95.7%. Wu et al. (2012) compared different clinical natural language processing systems on handling abbreviations in discharge summaries, resulting in MedLEE performing best with an F-Score of 0.60. A prototypical system, meeting real-time constraints, is described in Wu et al. (2013).

[8] http://www.nlm.nih.gov/research/umls/

## 5 Conclusion and Outlook

We have presented and evaluated a method for disambiguating the period character in German-language medical narratives. It is a combination of a simple rule set and a statistical approach supported by lexicons. Whereas the crafting of the rule base considers peculiarities of the document language, primarily by exploiting language-specific capitalization rules, the processing of the external language resources and the statistical methodology are unsupervised. Given these parameters, the accuracy values of about 93% for both abbreviation detection and sentence delineation are satisfactory, especially when one considers that the texts are error laden and highly compact, which also resulted in large numbers of ad-hoc abbreviations. We expect that with a limited training effort this rate can still be raised further. We are aware that the described period disambiguation procedure should be embedded into an NLP processing pipeline, where it must be preceded by a cleansing process that identifies “hidden” periods and restores the adherence to basic punctuation rules by inserting white spaces where necessary. An improved result can facilitate the creation of a sufficiently large, manually annotated corpus, which could then be used as the basis for the application of machine learning methods. Furthermore, the impact of the different modifications regarding the left context approach must be evaluated in more detail.

## References


Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

T Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

George Hripcsak and Daniel F Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of biomedical informatics*, 35(2):99–110.

T Kiss and J Strunk. 2002. Scaled log likelihood ratios for the detection of abbreviations in text corpora. In *Proceedings of the 19th International Conference on Computational Linguistics – Volume 2*, pages 1–5. Association for Computational Linguistics.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

S M Meystre, GK Savova, KC Kipper-Schuler, and JF Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 35:128–144.

JD Patrick, L Safari, and Y Ou. 2013. ShaARe/CLEF eHealth 2013 Normalization of Acronyms/Abbreviation Challenge. In *CLEF 2013 Evaluation Labs and Workshop Abstracts - Working Notes*.

Pschyrembel. 1997. Klinisches Wörterbuch. CD-ROM Version 1/97.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer.

Y Wu, ST Rosenbloom, JC Denny, A Miller, S Mani, Giuse DA, and H Xu. 2011. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, pages 1541–1549.

Y Wu, JC Denny, ST Rosenbloom, RA Miller, DA Giuse, and H Xu. 2012. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 997–1003.

Y Wu, JC Denny, ST Rosenbloom, Randolph A Miller, Dario A Giuse, Min Song, and Hua Xu. 2013. A prototype application for real-time recognition and disambiguation of clinical abbreviations. In *Proc. of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, pages 7–8.

H Xu, PD Stetson, and C Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 821–825.

H Xu, PD Stetson, and C Friedman. 2012. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 1004–1013.