Unsupervised Abbreviation Detection in Clinical Narratives

Markus Kreuzthaler[a,b]
[a]CBmed GmbH
Center for Biomarker Research in Medicine
Stiftingtalstrasse 5, 8010 Graz, Austria

Michel Oleynik[b], **Alexander Avian**[b] and **Stefan Schulz**[b]
[b]Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Auenbruggerplatz 2, 8036 Graz, Austria

Abstract

Clinical narratives in electronic health record systems are a rich resource of patient-based information. They constitute an ongoing challenge for natural language processing, due to their high compactness and abundance of short forms. German medical texts exhibit numerous ad-hoc abbreviations that terminate with a period character. The disambiguation of period characters is therefore an important task for sentence and abbreviation detection. This task is addressed by a combination of co-occurrence information of word types with trailing period characters, a large domain dictionary, and a simple rule engine, thus merging statistical and dictionary-based disambiguation strategies. An F-measure of 0.95 could be reached by using the unsupervised approach presented in this paper. The results are promising for a domain-independent abbreviation detection strategy, because our approach avoids retraining of models or use case specific feature engineering efforts required for supervised machine learning approaches.

1 Introduction

Free text narratives are a main carrier of unstructured patient-based information in clinical information systems. Clinical texts differ significantly from, e.g., newspaper or scientific articles. The following snippet demonstrates the high degree of compactness, which is typical for clinical narratives[1]:

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

As much as such highly condensed text is understandable by specialists, it poses severe problems to natural language processing (NLP) and subsequent semantic interpretation (Meystre et al., 2008), due to idiosyncrasies of telegram style language like word and term-level ambiguities, acronyms, abbreviations, single-word compounds, derivations, spelling variants and misspellings. In addition, the broad range of clinical specialties with different vocabularies and recording traditions account for a high variation of sub-language characteristics (Patterson et al., 2010).

This paper deals with the disambiguation of the period character (".") in clinical narratives. In many Western languages like German, periods are used as abbreviation markers. Therefore, in a first tokenization step it is not recommended to consider trailing period characters as token delimiters, in order to identify tokens that end with a period. Three cases can be distinguished: (i) The period character marks an abbreviation and does not act as sentence delimiter. (ii) The period character marks an abbreviation and also delimits the sentence. (iii) The period does not belong to the token and therefore delimits the sentence.

[1]English translation: "3. History of total excision of an exulcerated secondarily nodular superficially spreading melanoma (C43.5) of the outer left labia. Level 4, tumor diameter 2.42mm".

Proceedings of the Clinical Natural Language Processing Workshop,
pages 91–98, Osaka, Japan, December 11-17 2016.

Our approach is purely data-driven, which distinguishes it from recently published work (Wu et al., 2016; Griffis et al., 2016; Vo et al., 2016), predominantly based on supervised machine learning. In contrast, we avoid extensive manual annotations of training data as well as classification task triggered feature engineering, even though good results were obtained in a previous study (Kreuzthaler and Schulz, 2015). Another requirement is that the method should be easily adaptable to other clinical sub-language domains without model retraining or exhaustive dictionary or terminology management, and that classification results should be understandable in detail and traced back to core decision rules.

2 Materials and Methods

2.1 Data

Corpus: A sample of 1,696 de-identified German-language clinical in and outpatient discharge letters was obtained from the dermatology department of an Austrian university hospital. The documents were randomly assigned to a training and a test corpus, with 848 documents each.

Gold standard: From both corpora a list of word types followed by a period character was extracted by applying the following two regular expression sequentially:
(i) `\b\p{Graph}+\.(?=(\p{Punct}|\s|$))` matches any word type character sequence ending with a period character, and (ii) `([a-z]+\.|[A-Z][a-z]*\.)` filters the resulting types from step one by word characters without digits. About 2,300 word types ending with a period finally constitute the training and test set. Their content was manually annotated on whether the period character belongs to the word type or not. The inter-annotator agreement was very high, with a Cohen's kappa of 0.98 (Di Eugenio and Glass, 2004; Hripcsak and Heitjan, 2002).

Dictionary: An abbreviation-free medical dictionary (~1.45 million unique word types) was built using (i) a free contemporary German dictionary[2], (ii) a German medical dictionary (Pschyrembel, 1997), and (iii) texts from a consumer health Web portal[3]. All tokens ending with a period character were excluded from this resource, as a highly sensitive approach to keep it free of abbreviations. In addition, German abbreviations harvested from Web resources[4,5] (~5,800 acronym and abbreviation tokens) were excluded from the overall dictionary to make the final resource as abbreviation-free as possible, also accounting for potential punctuation errors in the three dictionaries such as missing abbreviation period markers. The resulting resource was used in our abbreviation detection strategy, as described in the following section.

2.2 Methods

Statistical approach: For the statistical classification approach we built a fourfold *observed* co-occurrence table $O(k_{nm})$ for every word type ending with a period character:

	Schema		*Example A*		*Example B*	
	Type	¬Type	"Pat"	¬"Pat"	"auf"	¬"auf"
•	k_{11}	k_{12}	300	17,970	8	18,262
¬•	k_{21}	k_{22}	78	66,718	1,322	65,474

Table 1: Two examples of *observed* corpus based frequency counts, *viz.* the two word types "Pat" and "auf", with and without a period as rightmost character (symbolized by •).

With the observed frequency counts $O(k_{nm})$ we calculate the log-likelihood ratio (LLR) (Dunning, 1993)[6,7] of a word type and its ending period character by use of Shannon's Entropy (Shannon, 1948):

[2] http://sourceforge.net/projects/germandict/
[3] http://www.netdoktor.at/
[4] http://de.wikipedia.org/wiki/Medizinische_Abkuerzungen
[5] https://de.wiktionary.org/wiki/Kategorie:Abkuerzung_(Deutsch)
[6] The Apache Mahout library was used for LLR calculation.
[7] http://tdunning.blogspot.co.at/2008/03/surprise-and-coincidence.html

$$H = -\sum_{i=1}^{n} p_i(log_b p_i) \quad (1)$$

$$LLR = 2 \cdot N \cdot (H_{matrix} - H_{rows} - H_{cols}) \quad (2)$$

For the cases mentioned in Table 1, LLR values amount to 579.11 for *Example A* and 571.56 for *Example B*. This has the advantage that per co-occurrence their relevance can be asserted assuming a χ^2 distribution (with one degree of freedom) for different significance levels. *Example A* and *Example B* have a very high LLR, which allows the conclusion that the occurrence of the word type left of the ending period character has a *significant* influence on the presence *or* absence of the final period character. In order to determine whether there is significant evidence *for the presence* or *for the absence* of the final period character we calculate, in a next step, the *expected* values $E(k_{nm})$ of the fourfold Table 1 via:

$$k_{11}^{Exp} = (k_{11} + k_{12}) \cdot (k_{11} + k_{21})/(k_{11} + k_{12} + k_{21} + k_{22}) \quad (3)$$

$$k_{12}^{Exp} = (k_{12} + k_{11}) \cdot (k_{12} + k_{22})/(k_{11} + k_{12} + k_{21} + k_{22}) \quad (4)$$

$$k_{21}^{Exp} = (k_{21} + k_{11}) \cdot (k_{21} + k_{22})/(k_{11} + k_{12} + k_{21} + k_{22}) \quad (5)$$

$$k_{22}^{Exp} = (k_{22} + k_{12}) \cdot (k_{22} + k_{21})/(k_{11} + k_{12} + k_{21} + k_{22}) \quad (6)$$

These equations lead to the following fourfold *expected* co-occurrence table $E(k_{nm})$:

	Schema		*Example A*		*Example B*	
	Type	¬Type	“Pat”	¬“Pat”	“auf”	¬“auf”
•	k_{11}^{Exp}	k_{12}^{Exp}	81	18,189	286	17,984
¬•	k_{21}^{Exp}	k_{22}^{Exp}	297	66,499	1,044	65,752

Table 2: Two examples of *expected* corpus based frequency counts, again with the word types “Pat” and “auf”, with and without a period as rightmost character (symbolized by •).

The final decision function is now straightforward, reconsidering the fact that the expected values $E(k_{nm})$ can be interpreted as the distribution within the table if there were no divergence from randomness: If $O(k_{11}) - E(k_{11}) > 0$ the period character belongs to the word type and marks an abbreviation, if $O(k_{11}) - E(k_{11}) \leq 0$ the period marker does not belong to it and can be interpreted as sentence delimiter. We apply this decision function regardless of the *LLR*-level of the token-period co-occurrences, but its influence is inspected in the *Combined approach* described below.

Dictionary approach

The dictionary-based approach for period character classification is done via a simple dictionary look-up of the token under inspection[8]. If the token (without trailing period) is found in the dictionary, we decide that it is not an abbreviation, otherwise the period character is considered as belonging to the token, which is therefore classified as an abbreviation. This strategy requires an abbreviation-free dictionary, as described in Section 2.1.

Combined approach

Our decision function in the combined approach is motivated by the fact that the tokens ending with a period have a distribution pattern as depicted in Figure 1. This has a fundamental influence on our decision function: (i) For a certain proportion of the token-period co-occurrences the statistical approach will have enough frequency information to give valid classification results, (ii) *but* there is a relevant long

[8]Due to the large number of about 1.45 million dictionary entries, we used an Apache Lucene index, cf. https://lucene.apache.org/core/

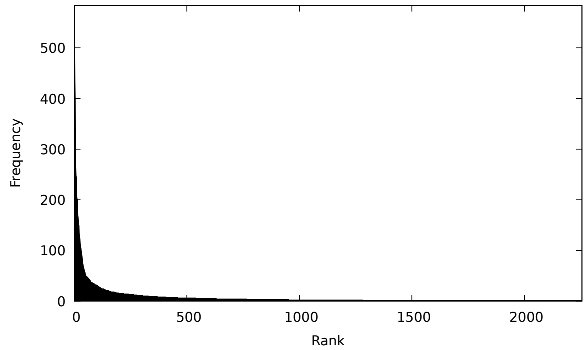

Figure 1: Ranked frequency count of tokens that end with a period.

tail of co-occurrences where the statistical method is not stable any more. We therefore addressed these cases by the dictionary-based approach and to prioritize it in the decision function: wherever the left context of the period is in the dictionary we decide in favor of a non-abbreviation, otherwise we take the decision of the statistical approach taking into account different significance levels ($LLR_1 > 10.83, p < 0.001$; $LLR_2 > 3.84, p < 0.05$; $LLR_3 > 0, p$-value not considered).

if *token* $\exists$ *dictionary* **then**
  $\rightarrow$ abbr=false;
else if $LLR >$ *significance level* **then**
  if $O(k_{11}) - E(k_{11}) > 0$ **then**
    $\rightarrow$ abbr=true;
  else
    $\rightarrow$ abbr=false;
  end
else
  $\rightarrow$abbr=true
end

Algorithm 1: Combined decision algorithm.

3 Results and Discussion

The evaluation results show that the *Statistical approach* on its own tends to find all abbreviations but lacks precision. The *Dictionary approach* returns an F-measure of 0.94, and the top performance result of $F_1 = 0.95$ is obtained with the combined approach. The evaluation results of the *Combined approach* also reflect the fact that the LLR information can be neglected in that case and the outcome of $O(k_{11}) - E(k_{11})$ should always be used regardless of the impact of the significance of the token-period co-occurrence. The investigation of false positives shows, e.g., a noticeable amount of token-period co-occurrences like “Lymphknotenstatus.” (in English “lymph node status.”) which very commonly appear at the end of a sentence, but which are not in our dictionary (search term: “Lymphknotenstatus”), and have a $O(k_{11}) - E(k_{11}) > 0$. False negative results typically appear with abbreviated tokens, such as “morph.” (abbreviation for ”morphologisch”, in English ”morphological”), which are erroneously found in our dictionary (search term: “morph”) and are therefore classified as non-abbreviations.

Kiss and Strunk (2002a), tried to reduce the amount of false positives and false negatives by applying different scaling factors to the resulting LLR. A final threshold was manually chosen, with F-measures of 0.92 and higher on newspaper corpora. Kiss and Strunk (2002b) performed an intermediate evaluation of their idea of re-scaling the LLR also for sentence boundary detection. Here, they obtained a minimum F-measure of 0.91. Both preliminary approaches finally led to the *Punkt* system (Kiss and Strunk, 2006), a multilingual unsupervised approach rigorously tested and evaluated. Kreuzthaler and Schulz (2014) applied an extended version of the Kiss and Strunk (2002a) method in an initial experiment with clinical

	Training			*Test*		
Method	Precision	Recall	F_1	Precision	Recall	F_1
Statistical						
Token	0.47	1.00	0.64	0.44	1.00	0.61
Type	0.36	0.97	0.53	0.35	0.97	0.51
Dictionary						
Token	0.90	0.98	0.94	0.88	0.97	0.93
Type	0.57	0.81	0.67	0.54	0.80	0.65
Combined$_{LLR_1}$						
Token	0.91	0.98	0.94	0.89	0.97	0.93
Type	0.57	0.81	0.67	0.56	0.80	0.66
Combined$_{LLR_2}$						
Token	0.92	0.97	0.94	0.90	0.97	0.94
Type	0.58	0.80	0.67	0.56	0.79	0.66
Combined$_{LLR_3}$						
Token	0.94	0.97	**0.95**	0.92	0.97	**0.94**
Type	0.61	0.78	0.69	0.59	0.78	0.67

Table 3: Evaluation results.

texts and achieved an accuracy of 0.93 for abbreviation and sentence detection based on the interpretation of the period character. A supervised machine learning approach using a support vector machine with a linear kernel and thorough feature engineering led to an F-measure of 0.95 for abbreviation detection and an F-measure 0.94 for sentence delineation (Kreuzthaler and Schulz, 2015).

Studies have also focused on the detection, normalization, and context-dependent mapping of abbreviations/acronyms to long forms (Xu et al., 2012). This is also part of works such as CLEF 2013 (Suominen et al., 2013), which included a task for acronym/abbreviation normalization, using the UMLS[9] 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). A decision tree classifier, which additionally used features from knowledge resources, performed best with a precision of 0.91 and a recall of 0.80. 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. A random forest classifier performed best with an F-measure of 0.95 and an ensemble of classifiers achieved the highest F-measure of 0.96. Wu et al. (2012) compared different clinical natural language processing systems for abbreviation handling in clinical narratives: MedLEE (Friedman et al., 1995b; Friedman et al., 1995a) performed best with an F-Measure of 0.60. A prototypical system, meeting real-time constraints, is described in Wu et al. (2013). Wu's journey finally ended in the CARD system (Wu et al., 2016) achieving an F-measure of 0.76 for finding *and* disambiguating abbreviations in clinical narratives. Very recently Vo et al. (2016) got very high results with a minimum F-measure of 0.94 on abbreviation detection on clinical notes applying supervised machine learning methods which a rich feature engineering process.

The main difference between the work we presented and the unsupervised approach of Kiss and Strunk is the fact that we refrained from re-scaling the LLR and avoided to set an experimental threshold for the abbreviation classification task. The statistical decision function we employed proved to be solid and robust even in cases where $k_{21} > k_{11}$ (e.g. "Meta." with $k_{11} = 28$, $k_{21} = 82$, but nevertheless correctly classified as abbreviation), which had also been one type of motivation for introducing scaling factors by Kiss and Strunk (2006). In contrast to much of the related work, our approach is unsupervised

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

and does not require the training of a machine learning model or a rich feature engineering effort (Vo et al., 2016; Wu et al., 2016; Kreuzthaler and Schulz, 2015). Therefore we hypothesize that our approach is especially suited to be deployed to other clinical domains, which was a main driver of our investigations. Table 3 shows that with the dictionary approach alone we got F-measure values greater than 0.93, whereas the performance by word types was much lower. For the time being, we consider this acceptable because we concentrate on high token-based evaluation measurements and do not want to misclassify frequently occurring abbreviations. The statistical approach is not applicable in isolation, because we have found many cases where a word type followed by a period occurs only once or twice in the corpus (see Figure 1). In such cases the statistical approach is not robust any longer, so we have to rely on dictionaries. The combined approach was satisfactory as both training and test yielded token-based F-measure values for period character disambiguation greater than 0.94.

4 Conclusion and Outlook

In this paper we presented an unsupervised approach for period character disambiguation in German clinical narratives, which we evaluated for the task of abbreviation detection. We motivated and introduced both a data-driven statistical approach and a dictionary-based method. Based on the analysis of the frequency distribution of token-period character co-occurrences we also presented a hybrid methodology. This hybrid approach put emphasis on the dictionary-based method, which was then supported by a statistical decision rule. A dermatology corpus was used for initial evaluation. For the training and test set, we obtained F-measures of 0.95 and 0.94, respectively. This supports the hypothesis that unsupervised approaches are well suited for abbreviation and sentence boundary detection in clinical narratives, which are known to abound with ad-hoc abbreviations. Furthermore, the system presented here needs no adjustment to the sublanguage, which makes it easy to reuse for other text genres and subject-matters. This consideration together with the ability to trace back decision results to their core classification logic and the avoidance of manual training data annotations were major drivers for this investigation.

We mention the following limitations: (i) Periods after digits are currently not considered despite their importance as markers of ordinals in many languages, as well as their importance in many data formats. Kreuzthaler and Schulz (2015) took this into account in a supervised rich feature engineering approach using support vector machines; (ii) The methodology presented in this paper cannot resolve cases where periods play a double role, *viz.* as both abbreviation markers and sentence delimiters. This can be addressed by including in-depth context information regarding the period character under investigation; (iii) We applied this method to only one kind of text, *viz.* medical discharge summaries of melanoma patients. Therefore, we plan to demonstrate domain independence by applying the same approach to cardiology reports; (iv) We only used German texts, so that we can say little about the generalizability to other languages. Although we have found that ad hoc abbreviation is a very common phenomenon also in other languages and text genres, it cannot always be taken for granted that the period character is used as a marker. Future investigations will address these problems. Our goal is to create a specific UIMA component for abbreviation detection and resolution with an unsupervised core, which could be integrated in a clinical NLP pipeline like cTAKES (Savova et al., 2010), The Leo framework - The VINCI-developed NLP infrastructure (Meystre et al., 2008; Patterson et al., 2014) or MedKATp.

Acknowledgements

This work has been carried out as part of the IICCAB project (Innovative Use of Information for Clinical Care and Biomarker Research) within the K1 COMET Competence Center CBmed, which is funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT); the Austrian Federal Ministry of Science, Research and Economy (BMWFW); the Austrian state of Styria (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency.

References

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

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

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995a. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(01):83–108.

Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. 1995b. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 347–351. American Medical Informatics Association.

Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M Lai. 2016. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. *AMIA Summits on Translational Science Proceedings*, 2016:88–97.

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

Tibor Kiss and Jan Strunk. 2002a. 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.

Tibor Kiss and Jan Strunk. 2002b. Viewing sentence boundary detection as collocation identification. *Proceedings of KONVENS 2002*, pages 75–82.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Markus Kreuzthaler and Stefan Schulz. 2014. Disambiguation of period characters in clinical narratives. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@EACL*, volume 100, pages 96–100.

Markus Kreuzthaler and Stefan Schulz. 2015. Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Medical Informatics and Decision Making*, 15(2):1–13.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Jon D Patrick, Leila Safari, and Ying Ou. 2013. ShARe/CLEF eHealth 2013 Normalization of Acronyms/Abbreviations Challenge. In *CLEF 2013 Evaluation Labs and Workshop Abstracts - Working Notes*.

Olga Patterson, Sean Igo, and John F Hurdle. 2010. Automatic acquisition of sublanguage semantic schema: Towards the word sense disambiguation of clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2010, pages 612–616. American Medical Informatics Association.

OV Patterson, TB Forbush, SD Saini, SE Moser, and SL DuVall. 2014. Classifying the indication for colonoscopy procedures: A comparison of NLP approaches in a diverse national healthcare system. *Studies in Health Technology and Informatics*, 216:614–618.

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

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

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 *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Thi Ngoc Chau Vo, Tru Hoang Cao, and Tu Bao Ho. 2016. Abbreviation identification in clinical notes with level-wise feature engineering and supervised learning. In *Pacific Rim Knowledge Acquisition Workshop*, pages 3–17. Springer.

Yonghui Wu, S Trent Rosenbloom, Joshua C Denny, Randolph A Miller, Subramani Mani, Dario A Giuse, and Hua Xu. 2011. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, pages 1541–1549.

Yonghui Wu, Joshua C Denny, Samuel Rosenbloom, Randolph A Miller, Dario A Giuse, and Hua 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.

Yonghui Wu, Joshua Denny, S Trent 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 *Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, pages 7–8.

Yonghui Wu, Joshua C Denny, S Trent Rosenbloom, Randolph A Miller, Dario A Giuse, Lulu Wang, Carmelo Blanquicett, Ergin Soysal, Jun Xu, and Hua Xu. 2016. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). *Journal of the American Medical Informatics Association*, page ocw109.

Hua Xu, Peter D Stetson, and Carol Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 821–825.

Hua Xu, Peter D Stetson, and Carol 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.