Token-level noise in large Web corpora and non-destructive normalization for linguistic applications

Felix Bildhauer[*]
FU Berlin/SFB 632
felix.bildhauer
@fu-berlin.de

Roland Schäfer[*]
FU Berlin
roland.schaefer
@fu-berlin.de

1 Introduction

We discuss token-level noise in Web corpora, using our own COW2012 corpora (Schäfer and Bildhauer 2012), which are up to 9 billion tokens large, as examples. Web corpora pose unique problems by being (at least potentially) very large and noisy at the same time. We show that it is partly because of this noise that they open up completely new roads in linguistic research. What can be called "noise" from a POS tagging perspective, might be considered valuable evidence in research on non-standard writing. Normalization therefore has to be implemented very carefully or non-destructively. To illustrate, we describe two types of normalization tools which we have developed as well as our improved architecture to make huge non-destructively normalized corpora available to linguists.

2 Token-level noise in DECOW2012

As an example of a huge and noisy corpus, we use our DECOW2012 corpus of German (9.1 billion tokens) which was compiled and processed from Web data in 2012.[1] Similar to the findings in Liu and Curran (2006) for other Web corpora, we have 63,569,767 different types (after tokenization), of which 39,988,127 are hapax legomena. Such figures are implausible for any single natural language, and there must be a huge number of "noisy tokens" from diverse sources in the corpus. We have assessed the sources and types of noisy tokens by extracting a sample of tokens which were unknown to the standard German model for TreeTagger (Schmid 1995). We found the distribution shown in Table 1. It is known that the usual proportion of over 50% of truly noisy tokens cause further damage (i.e., degraded accuracy) in the whole linguistic post-processing chain. A good example is the significantly lower POS tagger accuracy on Web data (e.g., Giesbrecht and Evert 2009, summary in Schäfer and Bildhauer 2013). While it is possible to extend a tagger's lexicon (or train a whole new tagger model) to improve the accuracy on the 46.8% of rare but real words and the 1.2% of numbers, dealing with the true noisy tokens involves a whole series of challenges and, above all, design decisions. We discuss two components of our normalization chain: one which applies destructive normalization (dehyphenation) and one which applies non-destructive normalization (spelling correction).

Source	%	±% CI
misspelling	20.0	5.0
tokenizer error	17.6	4.7
non-word	7.6	3.3
foreign language	6.8	3.1
real rare word	46.8	6.2
number	1.2	1.3

Table1: Sources of tokens which are unknown to TreeTagger in DECOW2012 (with 95% confidence interval for the estimate); true noisy tokens are above the line; below the line are those for which the POS tagger should (in principle) have a solution

3 Two examples of Web corpus noise

Hard-coded hyphenation is not only found in OCR'ed documents, but also in Web corpora. Sources include, e.g., texts from word processors or PDFs pasted into content management systems. The number of noisy tokens from hyphenation in Web corpora is relatively small. In Table 1, hyphenated words are included in the 7.6% of non-words, which also include HTML markup and similar material. In a language like German, automatic normalization of hyphenated words must distinguish between (1) ordinary hyphenation words like *Seiten- streifen* ("*hard shoulder*") which have to be concatenated while dropping the hyphen (→ *Seitenstreifen*), (2) compounds which are actually written with a hyphen like *Philipps- Lagerverkauf* ("*direct sale by Philips*") and were just accidentally ripped in halves at the hyphen position (→ *Philipps-Lagerverkauf*), and (3) abbreviated and coordinated compounds, which must never be normalized since they represent standard orthography: *weder TV- noch Radiosender* (*"neither TV nor radio stations"*).

First, it must be noted that simple approaches which rely on line endings being marked in the corpus (e.g., Grefenstette and Tapanainen 1994) will not work because the original line endings at which hyphenation occurred are usually absent in Web corpora. Our solution is HyDRA, an efficient (compiled) tool/library. It works in an unsupervised manner by generating frequency lists of unigrams and bigrams from a corpus in the training phase. In the production phase, it uses the database of ngrams and their frequencies as a primitive language model

[*] Both authors have made equal contributions.
[1] http://www.corporafromtheweb.org/

to decide which candidates (= bigrams where the first token ends in a hyphen) need to be treated as in example (1) or (2), and which need to be left alone as in example (3). With this approach, we achieve a token accuracy of 99.6%, but a type accuracy of only 63.7% on our German data set of 1,000 candidates. To improve the type accuracy, the system allows the additional definition of hand-crafted rules to override the decision based on the language model. A single such rule for German boosts the token accuracy to 99.9% and the type accuracy to 91.2%. This normalization step is applied destructively, i.e., corpus users cannot see the original source if HyDRA has dehyphenated a word.

Our second example of normalization deals with the more massive amount of misspellings among the noisy tokens (20%). In Web corpora of the given size, we can easily find over 20 misspelled variants of any ordinary lexical word. Misspellings include typos (*überninmmt* for *übernimmt "takes over"*; Levenshtein distance 1), lack of orthographic knowledge or dialectal spellings (*überniehmt*; Levenshtein 2) or even intentional "emphatic" spellings (*überniiiiiiiimmmmmmmt*; Levenshtein 12). We use Enchant[2] to get suggestions from diverse spell-checking libraries. Since the primary goal of normalizing spelling (for us) is improving POS tagger accuracy, it makes sense to use only suggestions from spell-checkers (instead of using more advanced techniques as in the dehyphenation scenario, possibly deriving suggestions from language models based on the corpus itself): Words known to spell-checkers are likely to be also included in the POS tagger's lexicon. Although tools like Aspell are quite good at suggesting the correct form even for cases like emphatic spellings, we do not destroy the original spellings, but merely add the corrected spelling as an additional annotation layer. (Tokens for which no correction was necessary are simply copied to the annotation layer.) The new layer is separately POS tagged and lemmatized, such that users can perform full searches within corrected and original word forms, POS tags and lemmas, and even across original and normalized layers. Trivially, this increases the amount of data by a factor of 2. Why we think this is necessary for corpora which are quite huge to begin with is explained in the next section.

4 Reasons for keeping the noise

We argue that most of what looks like noise in Web corpora must not be "normalized away", because many kinds of "noise" in fact offer entirely new possibilities in linguistic research on non-standard language and non-standard writing. From a POS tagging (or other post-processing) perspective, we absolutely need the normalization in order to assign correct POS tags and lemmas to tokens, but otherwise the data should be left intact. It is therefore that we are convinced that normalization should be applied in a non-destructive fashion whenever possible.

As an example, Schäfer and Sayatz (2013) show that certain cliticized variants of the indefinite article in German (full form *ein*, *einen*, *einem*, *eine* etc.), which are in frequent use in spoken language, are written in large quantities in DECOW2012 (as *n*, *nen*, *nem*, *ne* etc.). They do not occur at all in corpora of standard written language (like newspaper corpora). Although the phenomenon has been discussed in the literature, the aforementioned authors are the first to back up their claims by providing huge amounts of data (roughly 3.5 million cases) and statistical analysis. The contextual and morpho-phonological conditions which favor the selection of shortened forms can be stated clearly. Most importantly, the authors show that there is evidence for autonomic graphemic principles (involving graphemic syllables and graphemic words) which are independent of the grammar of the language. For example, full written cliticization (formation of a single graphemic word) as in *und ein* → *undn* or *und einem* → **undnem* (*"and a"*, masculine nominative and dative forms) is only possible when the form of the cliticized article is not an acceptable graphemic word by itself (like the single nasal grapheme *n* from *ein*), as opposed to a well-formed graphemic word (containing at least one vowel letters, like *nem*). If the reduced article constitutes an acceptable graphemic word, it is always written as a separate word (*und nem* or *und 'nem*), while otherwise it tends to form a single word with the previous word (*undn*). This kind of evidence can neither be found in corpora of standard written language nor in corpora of spoken language. The phenomenon is characteristic of and limited to non-standard writing as found in some Web genres. There are many other areas for which similar arguments are possible, for example research on spelling errors.

Since traditional, hand-crafted corpora will often be superior to Web corpora when it comes to research on standard language and standard writing, normalizing Web corpora destructively and too aggressively would thus seriously weaken one of their major their selling points.

5 Query architectures

To make the carefully normalized corpora available to linguists, an indexing and query engine has to be chosen, and the choice significantly influences the way we can apply and mark normalizations in the

[2] http://abisource.com/projects/enchant/

final product. We use the IMS Open Corpus Workbench (CWB)[3], because it allows us to index our corpora in quite large chunks (of roughly 1.5 billion tokens) without major performance loss. For corpora of the size under discussion in this paper, we know no alternative. CWB has drawbacks, however, when it comes to non-destructive normalization. For example, multi-token units (MTU) and multi-unit tokens (MUT) cannot be represented adequately in CWB. To dehyphenate non-destructively, MUT capabilities would be required, and to render tokens like *undn* (cf. Section 4) in a POS tagger-friendly fashion, MTU capabilities would be required. In other words, transparently mapping *Seiten- streifen* to *Seitenstreifen* or *undn* to *und n* (with unequivocal tags for each token) in CWB is virtually impossible. Other architectures, like the ANNIS tool (Zeldes et al. 2009), which are suitable to deal with this kind of annotation, fail to perform well (or rather: at all) with corpora in the giga-token region.

We decided to use destructive normalization for dehyphenation with HyDRA because of the high accuracy of the tool and the low frequency of the phenomenon. Also, hard-coded hyphenation (mostly created by software) seems to us to be true noise, also from the linguistic perspective. On the problem with *undn*, cf. Section 6.

The increased size caused by normalization layers in the new version of our corpora (e.g., DECOW2013, scheduled for release in July 2013) makes it necessary to split the corpora in even smaller slices (roughly 20 for DECOW2013). This means that querying the whole corpus becomes quite cumbersome for linguists who cannot write their own scripts and cannot use parallelization. Around 20 single queries would have to be performed and merged manually for each lookup. Therefore, we are currently developing a simplistic Map-Reduce abstraction layer for CQP which executes CQP child processes on multiple machines. The tool does not require setting up a whole complicated cluster infrastructure (like Apache Hadoop). Processes across any number of machines communicate via a simple SSH connection and common NFS file systems. A Reduce is currently available for concordances, and we are working on a Reduce for CWB group results (combinatorial frequency tables).

6 Open problems and outlook

Some problems remain unsolved for the time being. To name just one example, POS tagging non-standard cliticized words like the aforementioned *undn* (*"and a"*) is impossible with most available POS tagger models. We are convinced that new tag sets and tagger models need to be developed which accomplish the task of assigning some reasonable single POS tag to such forms and map them to the lemma of the non-clitic element (in the example, this is *und*). We hope to present first evaluations of our work on POS tagger improvement at the workshop, at least for English and German.

We also hope to present an evaluation of HyDRA and the spelling correction for our other Web-derived corpora (currently Danish, Dutch, English, French, Swedish). If there is significant interest, HyDRA can be released as a cross-platform library with a C ABI and C header files.

[3] http://cwb.sourceforge.net/

References

Giesbrecht, E. and Evert, S. 2009. "Part-of-Speech (POS) Tagging -- a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus". In: Alegria, I. and Leturia, I. and Sharoff, S. (eds.) *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*: 27-35

Grefenstette, G. and Tapanainen, P. 1994. "What is a word? What is a sentence?" In: *Proceedings of 3rd Conference on Computational Lexicography and Text Research*.

Liu, V. and Curran, J. R. 2006. "Web text corpus for natural language processing". In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. EACL: 233-240.

Schäfer, R. and Bildhauer, F. 2012. "Building Large Corpora from the Web Using a New Efficient Tool Chain". In Calzolari, N. and Choukri, K. and Declerck, T. and Doğan, M. U. and Maegaard, B. and Mariani, J. and Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. ELRA: 486-493.

Schäfer, R. and Bildhauer, F. 2013, to appear. *Web Corpus Construction*. San Francisco: Morgan and Claypool.

Schäfer, R. and Sayatz, U. 2013, submitted. "Die Kurzformen des Indefinitartikels im Deutschen".

Schmid. H. 1995. "Improvements in Part-of-Speech Tagging with an Application to German". In: *Proceedings of the EACL SIGDAT-Workshop*, Dublin, Ireland.

Subramaniam, L. V. and Roy, S. and Faruquie, T. A. and Negi, S. 2009. "A survey of types of text noise and techniques to handle noisy text". In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, ACM: 115-122.

Zeldes, A. and Ritz, J. and Lüdeling, A. and Chiarcos, Chr. 2009. ANNIS: "A Search Tool for Multi-Layer Annotated Corpora". In: *Proceedings of Corpus Linguistics 2009*.