Automatic Detection of Linguistic Quality Violations

Jonathan Oberländer

20th June 2014

1 Corpus Analysis

The first step is to inspect the corpus. It is soon noticeable that the type system of the LQVCorpus (Valeeva, 2013) isn't detailed enough for our purposes and has a few shortcomings: Especially under the label of `other_ungrammatical_form` many different types of errors are combined. For this reason, we define a number of subtypes for this violation:

missing spaces

Comment: rename to missing spaces? But there are cases (which?) where this does not apply.

The most common type of error. The sentence contains a word that does not exist. In almost all cases, this happens when whitespace between two words is missing.

Example: *A strong earthquake measuring 7.8 magnitude struck **Wenchuancounty** of Sichuan Province on Monday, leaving at least **12,000people** died and thousands more injured.*

missing words

One or multiple words are clearly missing. These seem to most often be function words such as articles or pronouns, rather than content words. In the example, an underscore marks the position where a word, probably "was" is missing.

Example: *An Israeli woman _ killed and 11 others were wounded in the suicide bombing at a shopping mall in southern Israeli town of Dimona.*

punctuation error

Most of the time, this means: Punctuation is missing, but it can also mean there is something else wrong with punctuation in that sentence which makes it ungrammatical, as can be seen in the example: Unbalanced parantheses.

Example: *China has allocated 200 million yuan (million dollars for disaster relief work after an earthquake rocked the country's killing at least seven people, state reported on Tuesday.*

capitalisation error

A word that should be capitalised isn't or one that shouldn't be capitalized is.

Example: *earlier on **m**onday **g**erman chancellor **a**ngela **m**erkel and foreign minister **f**rank **w**alter **s**teinmeier offered their condolences to **c**hina over the heavy loss of life in the powerful earthquake that hit **c**hina's southwestern province of **s**ichuan.*

ungrammatical/unparsable

Comment: Unsure about the naming

This subtype looks similar to a punctuation error, but differs in that sentences are intermixed with each other; in the middle of one sentence, the reader suddenly finds themself in a different one. This could also happen if part of a sentence was removed. In the example, an underscore marks the point at which the break happens.

Comment: wording

Example: *All of those provinces and Chongqing, a special municipality _ deepest condolences to those who lost their loved ones in the devastating natural disaster.*

heading

The sentence contain (usually at the beginning) a sequence of capitalised words that aren't part of a dateline.

Example: *THE CURRENT FIX: Internet applications such as firewalls and spam filters attempt to control security threats.*

incomplete sentence

The type system of the LQVCorpus (Valeeva, 2013) defines an `incomplete_sentence` violation as words being cut off at the end of a sentence. A couple of times though, this also occurs at the beginning of a sentence. We restructure the type system by treating all types of incomplete sentences as this subtype of `other_ungrammatical_form`.

Example: *A Palestinian suicide bomber detonated an explosive belt at a commercial center in Dimona on.*

not ungrammatical

As is to be expected, the annotation of the corpus isn't correct 100% of the time. This subtype merely denotes a correct sentence that was incorrectly marked as ungrammatical.

Example: *One Israeli woman was killed and at least eight others wounded on Monday in a suicide bombing which ripped through a commercial center in the southern Israeli town of Dimona, the first attack of the kind since January 2007.*

The distribution of these types in *dev-1*, the first 20% of the LQVCorpus, can be seen in Figure 1.

We then mark these types in *dev-2* for every sentence that is tagged as `other_ungrammatical_form` in the corpus.

Perhaps the most noticeable result is the large portion of *missing spaces* violations, meaning clauses that include tokens which aren't correct words. Almost all

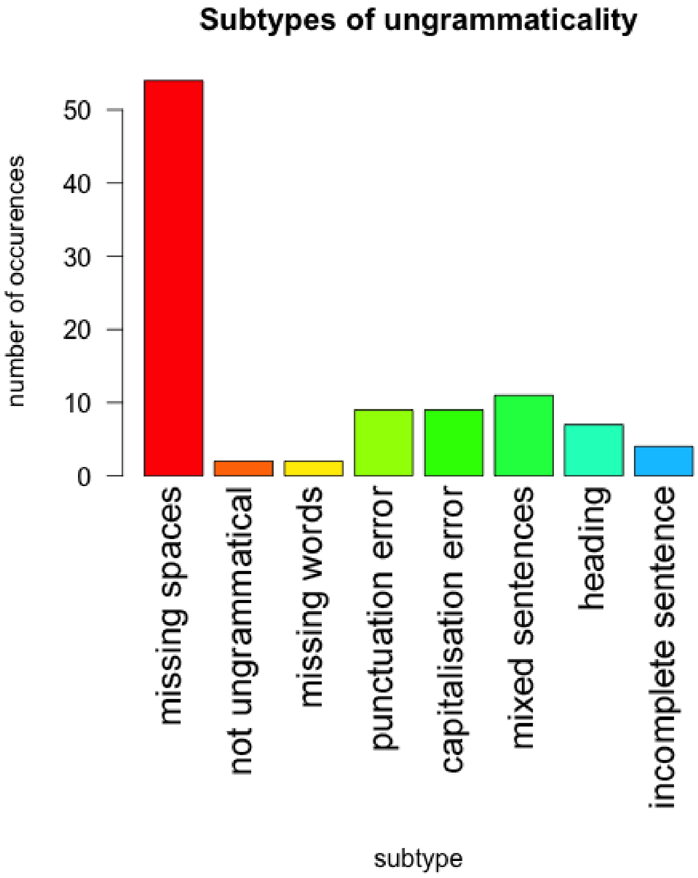

Figure 1: **Subtypes of ungrammaticality in *dev-1*.**

of these cases came from a missing space between two words forming tokens such as `reportsreaching` or `Wenchuancounty`. As there is such a large amount of this type, finding a reliable detection method for this subtype would significantly boost detection of ungrammaticality in general.

2 Method

The annotation, our work and the evaluation are done on sentence basis. As the text isn't split into sentences in the raw corpus, we use Stanford CoreNLP for that task.

Being the biggest amount of cases of ungrammaticality, *missing spaces* violations seem to have a straightforward solution which we call the **UnknownTokens** approach: Tagging a sentence as containing a *missing spaces* ungrammaticality if there is a token that isn't a "known" token, i.e. one that doesn't exist in a list of known words. Obviously, this doesn't include words for named entities yet, so as a further condition they should also not be tagged as such by a Named Entity Recognizer.

The Stanford Named Entity Recognizer (Finkel et al., 2005) is a widely used, state-of-the-art NER that comes with a model for the English language. After it was set up with the toolchain that is used to interact with the LQVCorpus XML, a list of known tokens was generated from the first 20% of the AFE part of the gigaword corpus (Graff et al., 2003) and violations of the type *Other ungrammatical form* in *dev-2* (the second 20% of the LQVCorpus) were annotated with the corresponding subtype.

In order to increase recall for NER, we assume that words that start with a capital letter and are only followed by lower case letters are known words / named entities. Finally, we automatically check whether an unseen token has a wikipedia entry, which further improves precision. We automatically label any tokens that are neither on our list of known tokens from GigaWord, nor tagged as a named entity as "unknown".

3 Evaluation

We evaluate our experiments using the well-known metrics **P**recision, **R**ecall, **F**-Score and **A**ccuracy:

$$P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn}$$

$$F = 2 * \frac{P * R}{P + R} \quad A = \frac{tp + tn}{tp + tn + fp + fn}$$

For experiment 1, we test how well our system works for detecting missing spaces. All sentences in `dev-2` that are correctly classified as `other ungrammatical form` with the subtype `missing spaces` are seen as true positives (tp), the ones it missed as false negatives(fn). Sentences wrongly tagged as containing missing spaces are counted towards false positives(fp) and finally, everything else is a true negative (tn).

	Precision	Recall	Accuracy	F-Score
Experiment 1 (missing spaces)	42.85%	100.00%	95.03%	59.99%
Experiment 1 (no missing spaces)	96.05%	99.32%	95.65%	97.66%
Experiment 1+wiki (missing spaces)	78.04%	88.89%	98.14%	83.11%
Experiment 1+wiki (no missing spaces)	98.68%	99.34%	98.17%	99.01%
Baseline (missing spaces)	0%	0%	94.40%	0%
Baseline (no missing spaces)	94.40%	100%	94.40%	97.12%

Table 1: Evaluation of **UnknownTokens**

	Experiment 1 (missing spaces/no missing spaces)
Micro-average precision	92.77%
Macro-average precision	69.45%
Micro-average recall	98.72%
Macro-average recall	99.66%
Micro-average F-score	95.65%
Macro-average F-score	78.83%

Table 2: Macro- and micro-averages for **UnknownTokens**

In experiment 2, with the same data, we use the *missing spaces* type as a measure for ungrammaticality and thus consider a sentence a true positive if our system detects a nonword and the sentence is tagged as ungrammatical in the corpus.

A evaluation lead to the results shown in Table 2. The perhaps surprisingly high accuracy is due to the fact that the “correct” sentences outweigh the ungrammatical ones, leading to a high true negative count. This can also be seen in the baseline, which consists of tagging every word as not containing a *missing spaces* violation.

In a second set of experiments (1a and 2a, also shown in Table 2), we varied the settings of our previous experiments: In experiment 1a, we additionally annotated sentences of the violation type `incomplete sentence` with the subtype nonword, if applicable. In experiment 2a, we looked into how good the system could decide whether a sentence was either one of `incomplete sentence` or *other ungrammatical form*, which lead to an increased precision, along with a severe decline in recall, because most incomplete sentences don’t contain nonwords. It is to be expected that combining this approach with a seperate system for detection of cutoff sentences will lead to a much higher performance.

4 Discussion

When looking into the false positives, this result can be broken down to one main issue:

Clauses can contain multiple violations, such as *other ungrammatical form* and *incomplete sentence*. As only one of them was annotated, and the subtype was only added to clauses of the ungrammaticality type *other ungrammatical form*, a large portion of the false positives isn't actually detected incorrectly, but rather not fully annotated. The fragment

> A spokesman for al-Aqsa Martyrs Brigades, armed wing **ofPalestinian** Fatah movement, denied on Monday the reports that **thetwo** suicide

was annotated (correctly) as *incomplete sentence*, but it is clear that the sentence also contains general ungrammaticality. As such, the system detected it as an unknown token and was punished for that decision.

Adressed in exp 1a and 2a, rewrite

References

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Graff, D., Kong, J., Chen, K., and Maeda, K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*.

Valeeva, M. (2013). Annotation of factors of linguistic quality for multi-document summarization.