DIGITHUM
Les humanitats en l'era digital

www.uoc.edu/digithum

A Grammar Checker Based on Web Searching

Joaquim Moré
Researcher at the Internet Interdisciplinary Institute (IN3) of the UOC
jmore@uoc.edu

Submission date: January 2006
Published in: May 2006

Recommended citation:

MORÉ, Joaquim (2006). "A grammar checker based on web searching". *Digithum*, issue 8 [article online].
DOI: http://dx.doi.org/10.7238/d.v0i8.529

Abstract
This paper presents an English grammar and style checker for non-native English speakers. The main characteristic of this checker is the use of an Internet search engine. As the number of web pages written in English is immense, the system hypothesises that a piece of text not found on the Web is probably badly written. The system also hypothesises that the Web will provide examples of how the content of the text segment can be expressed in a grammatically correct and idiomatic way. Thus, when the checker warns the user about the odd nature of a text segment, the Internet engine searches for contexts that can help the user decide whether he/she should correct the segment or not. By means of a search engine, the checker also suggests use of other expressions that appear on the Web more often than the expression he/she actually wrote.

Keywords
grammar checking, style checking, natural language processing

Resum
En aquest article presentem un corrector gramatical de l'anglès destinat a escriptors no angloparlants. La principal característica d'aquest corrector és l'ús d'un motor de cerca per Internet. Com que hi ha un gran nombre de pàgines web escrites en anglès, el sistema fa la hipòtesi que un segment de text que no és present en cap pàgina web és probablement un segment de text mal escrit. El sistema també fa la hipòtesi que a la Xarxa hi trobarà exemples que ensenyaran a l'usuari com ha d'expressar el contingut del segment de text d'una manera gramatical i idiomàtica. Per tant, un cop el corrector avisa l'usuari que és millor verificar un segment del seu text, el motor cerca contextos que poden ser útils a la persona que escriu a l'hora de decidir si corregeix el segment o no. Gràcies també a l'ús d'un motor de cerca, el corrector suggereix a l'escriptor que utilitzi expressions que són més freqüents a la Xarxa en comptes de l'expressió que ha escrit.

Paraules clau
correcció gramatical, correcció estilística, processament del llenguatge natural

1. Introduction

The grammar and style checker presented herein is currently under development at the UOC. It is intended to help its researchers write texts in English, which is not their mother tongue. Although their command of the language is generally acceptable, most of them do not feel confident enough about the correctness and the idiomatic expressiveness of their writing. They feel confident about the correctness of a piece of text when they find it in a document already written in English (provided that this document is judged as grammatically and stylistically correct). However, if the piece of text is not found, the inference that it is probably badly written is only justified if the documents available are sufficiently numerous and varied. Internet provides an immense number of varied documents written in English; so the main characteristic of the checker is the use of an Internet search engine that detects the text segments that are not found on any web page. For each of these segments, the checker informs the user that the segment is brand new in the Internet universe and that it may be badly written, which is highly probable when the writer is not a native English speaker and does not have a sound knowledge of the language. The checker then searches for web pages containing different ways of expressing the content of the segment (variants). The search results page shows users contexts with the variants.

In the Natural Language Generation field, evidence from corpora has been used to choose a particular sentence construction (Langkilde & Knight, 1998; Langkilde, 2002) and Internet search engines have been used for testing error-detection rules in grammar checkers (Naber, 2003). This corpus-based checker never tells the user how to write. It would be against the creative use of language to judge a segment as 'incorrect' because it is not found on the Web. So the checker simply warns the writer and displays excerpts from the web pages found that contain variants of the segment written. These excerpts are considered helpful to the user for the detection of grammatical and stylistic errors, or in deciding whether to rewrite the text or not. Of course, the user can leave the segment as it is when the examples are not convincing enough for him/her to change it.

2. Description of the components

The checker has the following components:

- User Interface
- Tagger
- Chunker
- Internet search engines
- Brand new segment detector
- Improvable segment detector
- Searcher and displayer of examples

User Interface

The user interface loads the document the user wants to check (currently the document has to be in .txt format). The user can check a particular piece of text by clicking on it. In this case, the system checks the segment selected. If not, the system checks the whole text.

Tagger

The tagger annotates any string of words with their part-of-speech category (POS). The tagger used is the demo version of TreeTagger (Schmid, 1994) for Windows.[www1] The demo version cannot annotate more than 200 words. Nonetheless, the system currently focuses on the checking of segments selected by the user, so the number of words will hardly ever surpass this limit. The output of the tagger is a list of tagged words with the following format: Word-POS-Lemma.

Chunker

The chunker splits a POS-tagged piece of text into chunks. The chunks established so far are as follows:

- *Nominal*: string of words that are determiners, adjectives or nouns, and form an NP (e.g. *an Internet search engine*).
- *Verbal*: string of words that form single verbs and complex verbs.
- *Verbal+Nominal*: string of words containing a verbal followed by a nominal (e.g. *organise the academic activity*).
- *Nominal+Prep+Nominal*: string of words containing two nominals linked by a preposition (e.g. *labourer on a farm*).
- *Verbal+Prep+Nominal*: string of words containing a verbal and a nominal linked by a preposition (e.g. *carry out a project*).
- *Prep+Nominal*: string of words containing a preposition followed by a nominal. This string is not embedded in a larger chunk (e.g. *on the one hand*).
- *Adverbial+Verb/Adjective*: string of words containing an adverb and a verb or adjective (e.g. *also display examples*).

The chunks show concepts and relationships between the concepts in the segment. Prepositions and verbs are considered to be words that link concepts.

[www1]: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Search engines

The checker uses the online Wordnet 2.0[www2] engine, a lexical-semantic database, to get lexical information about how concepts can be expressed. The engines used to find the search results for a text segment are Yahoo[www3] and Altavista.[www4]

Brand new/improvable segment detectors

From the search results page, these detectors discern if the segment is brand new (no exact match found on any web page). If not, the detectors also judge if the segment can be improved (improvable).

Searcher and displayer of examples

When a segment is brand new or is judged to be improvable, this component searches for web pages containing variants of this segment and displays the snippets from the results page. These snippets may help the user reword the content of the segment. The maximum number of snippets that can be displayed on a search results page has been set at 100.

3. Detection of brand new and improvable segments

Brand new segments are those where the search results pages contain the sequence "We didn't find any web pages" where there are no snippets (out of 100) with the exact match highlighted. The detection of improvable segments is more complex.

3.1. Wordnet and the detection of improvable segments

The improvable segment detector activates the Wordnet search engine in order to find better wordings for a piece of text. For instance, when the syntactic chunk of the text segment is *Prep+Nominal*, the detector hypothesises that the piece of text is a way of expressing a concept, or a discourse connector. Due to the way Wordnet is organised, the engine searches for the nominal head's synsets (a synset is a set of synonyms denoting a concept) and the glosses that explain each synset. Each synset gloss containing the head in the results page is tagged and split into syntactic chunks. Then the chunk of the text segment is compared to the chunks that contain the head in the glosses. If the chunks coincide except for one non-functional word, then the text segment is regarded as improvable. Here is an example. Imagine the user wrote:

(1) *In the one hand, we explain the antecedents in the study of the cognitive processes…*

In the one hand is not brand new. But the Wordnet engine finds *on the one hand…, but on the other hand…* in the gloss for sense 7 of 'hand'. After tagging and chunking the gloss, the detector notices that *on the one hand* forms the same syntactic chunk as *in the one hand*, which does not appear anywhere in the Wordnet results page. So, the checker displays the following message:

(2) **hand** - (one of two sides of an issue; *on the one hand…, but on the other hand…*)

This message is the complete Wordnet information for sense 7 of 'hand'. This message may help the user realise that *in the one hand* should be revised.

3.2. Taking advantage of the "Did you mean…?"

When the question "Did you mean…?" appears in the search results page, the guessed form is tagged and split into chunks in order to check if the syntactic structure of the guessed form is the same as that in the text segment. If so, the guessed form is searched and the number of results compared to the number of results of the text segment. The text segment is regarded as improvable when the number of its results is smaller. For example, imagine the user wrote:

(3) *…it displays real English examples with an Internet searcher.*

The results page for *Internet searcher* contains the question "Did you mean 'Internet search'?" *Internet search* is tagged and identified as a noun phrase, as was *Internet searcher*. So, the results for *Internet searcher* (1,660) and *Internet search* (3,220,000) are compared. As a result of the comparison, *Internet searcher* is regarded as improvable.

3.3. Detecting the most frequent variant

A variant of a segment can be a string with the same words but in a different order. See, for example:

(4) *…in order to detect odd pieces of text and to also display helpful contexts.*

[www2]: http://www.cogsci.princeton.edu/~wn
[www3]: http://www.yahoo.com
[www4]: http://www.altavista.com

If the user wants to check *and to also display*, the adverbial 'also' is placed leftmost and then new queries are performed by moving the adverb one position each time from left-to-right. The engine searches for each variant and the detector compares the number of results (*also and to display*: 0; *and also to display*: 340; *and to also display*: 13; *and to display also*: 2). As the results for *and to also display* only exceed those for *also and to display* and those for *and to display also*, this segment is considered improvable.

4. Displaying helpful contexts

When a segment is considered improvable, the checker displays short excerpts from web pages containing the preferred variant. These contexts are the snippets from the results page. The variant appears in boldface type. So, in the case of *Internet search*, the system displays contexts such as (5i) and (5ii).

(5) i) …**Internet Search** *Tools. Single SearchEngines/ Portals…*
ii) *With billions of pages on the Web, you use a search engine if you're looking for something specific. Learn how search engines acquire, store and organize all that data to help you find what you're […] like most people, you visit an* **Internet search** *engine.*

After reading (5ii) the user who wrote *Internet searcher* may prefer to write *Internet search engine*. This is an example of how the system can be useful for translators, who have to handle terminology.

In the case of *to also display*, contexts like (6) are displayed:

(6) …*Sometimes the use of a spreadsheet can help the pupils to perform calculations more easily and* **also to display** *their results graphically in the form of bar charts and pie charts. This facility to…*

With respect to brand new segments, the search for helpful contexts is performed by substituting the words that link terms with a new element. When the segment is a *Verbal+Nominal* chunk, the verb is substituted by one of its synonyms. The synonym belongs to the synsets of the verb according to the results page from the Wordnet engine. Then the Yahoo and Altavista engines search for documents with the new keywords. If contexts are found, they are displayed to the user. For example, if the user writes the brand new segment *to devise the academic activity*, 'devise' is substituted by a different Wordnet synonym ('organise', 'organize', 'machinate'…) in *n* searches, where *n* is the number of elements in the verb's synsets. Then, contexts such as (7) are displayed to the user.

(7) *Committees including the important General/Professorial/ Academic Board, and the Finance Committee […] and lectureships, and* **organise the academic activity** *of specific departments or…*

If the synonym substitution fails, the words that link concepts (e.g. prepositions) are substituted by a special symbol that matches any word between the terms in question. The system displays the snippets from the results page where the terms are linked by a string of words in boldface (with no punctuation in between). In this string, the user can see a different preposition other than the one he/she used or learn an idiomatic way of linking the terms. The snippets are tagged and chunked in order to present first the contexts where the boldface words form the same syntactic chunk as in the original text segment. For example, if the user wrote *we carried up a project that lasted 2 years*, where *carried up a project* is brand new, the checker first displays contexts like *How we* **carried out our project** that may help the user realise that the preposition should have been 'out'.

A goal for the near future is to display contexts where certain terms of the original text that coexist in the sentence level (with no punctuation in-between) coexist in a more frequently used syntactic chunk. More idiomatic ways of saying the same thing would be presented to the user. For example, it would display **search results page** (an NP with 515,000 results) in the case where the user wrote *the page that shows the results of the search* (1 result). The system should consider this complex NP as a shorter way of stating the concept relations expressed in the sentence.

5. Comparison with other checkers

The checker presented here is different from the traditional ones in that it is not based on predefined language-dependent rules (Naber, 2003), tree-parsings (Jensen et al, 1993) or statistics (Atwell, Elliot, 1987). Except for the tagger, the other modules work using a search engine, which is "non-language-dependent". So the checker could easily be adapted to another language, provided a tagger for this language exists and can be called by the checker, and that the number of web pages in this language is large enough. Likewise, this checker could warn users about a wider range of phenomena beyond subject-verb agreement and other typical errors that are dealt with by the traditional systems. Indeed, this checker is being developed as complementary to these systems. Typical grammar and spelling mistakes are already detected by traditional checkers, so this is a simple way of assisting users whose writing skills involve aspects that are hard to detect by using predefined rules. As the system is currently under

development, evaluation data on its performance in order to compare with other checkers have not been obtained yet.

6. Future work

The first thing to do is to evaluate how the checker overcomes certain problems which are inherent in web searching. For example, badly written pages are not discriminated on the Web so the checker does not know for certain if a non-brand new segment matches the mistake of a non-native English writer. Case insensitive matching also causes some badly written segments to be considered as non-brand new. According to Naber (2003), Google finds the ungrammatical segment 'the is' because it matches a document containing *About the IS associates*, where 'IS' is probably an acronym.

Ungrammatical non-brand new segments are expected to be infrequent on the Web, but what is the minimum number of results necessary to judge a segment as grammatically correct? When the coexisting terms are very frequent, the threshold can be high (e.g. 'machine translation': 280,000 results) but the presence of a less frequent combination in a perfectly written segment leads to the number of results dropping dramatically (e.g. 'machine translation methods', 109 results); so the level should be set accordingly. Applying statistical methods could set the results threshold although other complementary methods are being considered, such as the identification of reliable URLs for the contexts displayed. For example, documents from URLs with *.edu* or containing *www.citeseer*, the huge online library of scientific publications, are probably written in an acceptable English.

Another problem inherent with search engines is their lack of linguistic criteria when matching. For instance, they do not match 'I loved the woman' with documents containing 'I love the women'. Queries to Wordnet and the tagging and chunking of snippets are expected to lessen these effects. This will be analysed and quantified in the near future.

References

ATWELL, E.; ELLIOT, S. (1987). "Dealing with ill-formed English text". In: *The Computational analysis of English*. Longman.

JENSEN, K.; HEIDRON, G.E.; RICHARDSON, S.D. (eds) (1993). *Natural language processing: the PLNP approach*. Kluwer Academic Publishers.

LANGKILDE, I.; KNIGHT, K. (1998). "Generation that exploits corpus-based statistical knowledge". In: *Proceedings COLING-ACL*.

LANGKILDE, I. (2002). "An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator". In: *Proceedings of the International Language Generation Conference 2002*. New York. P. 17-24.

NABER, D. (2003). *A Rule-Based Style and Grammar Checker*. Bielefeld University.

SCHMID, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94)*. Manchester, England. P. 44-49.

Joaquim Moré
Researcher at the Internet Interdisciplinary Institute (IN3) of the UOC
jmore@uoc.edu

He is a Researcher at the IN3, Technician at the UOC's Language Services, and specialist in language technologies. He is a graduate in English, and has a master's degree in Computational Linguistics from the University of Barcelona. He is currently working on his PhD thesis on the evaluation of machine translations.