14 – Automatic language correction

IA161 Advanced Techniques of Natural Language Processing

J. Švec

NLP Centre, FI MU, Brno

December 21, 2015

Motivation

This tool can be use to find spelling , gramar or stylistic errors in english texts. just paste some text in the the box and click 'Submit to check . Additionally, their are many different dialects you can chose from. Additionally , you can hover your mouse over a error to see it's description and an useful list of posible corrections. You don´t need to worry for your writing skills any more, improving you're text has never be more easier!

Figure: Example of text with errors.

Motivation

This tool can be use to find spelling , gramar or stylistic errors in english texts. just paste some text in the the box and click 'Submit to check . Additionally, their are many different dialects you can chose from. Additionally , you can hover your mouse over a error to see it's description and an useful list of posible corrections. You don´t need to worry for your writing skills any more, improving you're text has never be more easier!

Types of errors[1]:

Grammar (6) | Spelling (10) | Other (2) | Spacing (3) | Typographical (2) | Duplication (1)

[1] Source: `http://www.onlinecorrection.com/`

1. Spell checking
   - Type of errors
   - Error correction

2. Grammar checking
   - Rule-based grammar checking
   - Statistical grammar checking

3. Word completion

Automatic language correction

A text without errors...

- is easier to read,
- looks more professional,
- is easier to translate with machine.

Language correction has many areas:

- **spell checking** - detect spelling errors in individual words,
- **grammar checking** - incorrect use of person, number, case or gender, improper verb government, wrong word order, etc. . .
- **word completion** - suggestion of the word currently being entered.

Spell checking

- Detecting and providing spelling suggestions for incorrectly spelled words in a text,
- predicting which words in a document are misspelled,
- correction is the task of substituting the well-spelled hypotheses for misspellings,
- usually uses a dictionary of valid words,
- application: wordprocessing and postprocessing Optical Character Recognition [Whitelaw et al., 2009].

Type of errors

- **Non-word errors** - the misspelled word is not a valid word in a language,
  - typographic errors - usually keyboard mis-punches (e.g."teh" - "the", "speel" - "spell"),
  - cognitive errors - caused by the writer's misconceptions (e.g. "recieve" - "receive", "conspiricy" - "conspiracy"),
  - phonetic errors - substituting a phonetically equivalent sequence of letters (e.g."seperate" - "separate").
- **Real-word errors** - sentence contains a *valid* word, but it is *inappropriate* in the context [Hladek et al., 2013].

Example

Non-word error: "I'd like a peice of cake."
Real-word error: "I'd like a peace of cake."

Error correction

- Consists of two steps:
  - generation of candidate corrections,
  - ranking of candidate corrections.
- Isolated-word methods:
  - edit distance,
  - similarity keys,
  - n-gram-based techniques,
  - rule-based techniques,
  - probabilistic techniques,
  - neural networks [Gupta and Mathur, 2012].

Isolated-word methods I

- **Edit distance:**
  - assumption - person usually makes few errors,
  - minimum operations to transform a non-word to dictionary word,
  - operations: insertions, deletions and substitutions,
  - useful for: correcting errors resulting from keyboard input.

Example

Edit distance between "kitten" and "sitting" is 3:

1. kitten - sitten (substitution of "s" for "k"),
2. sitten - sittin (substitution of "i" for "e"),
3. sittin - sitting (insertion of "g" at the end).

Isolated-word methods II

- **Similarity keys:**
  - assign a key to each dictionary word,
  - compare with the key computed for the non word,
  - most similar key is selected as suggestion.
- Soundex – phonetic algorithm (English) [Holmes and McCabe, 2002].

Example

N	Represents letters
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

1. Retain the first letter.
2. Drop occurrences of a, e, i, o, u, y, h, w.
3. Replace letters with numbers.
4. Merge adjacent numbers.
5. Add zeroes to the end,

or remove rightmost numbers.
Output is (letter, number, number, number).

Robert - R163; Robin - R150 - not OK
Smith - S530; Smyth - S530 - OK

Isolated-word methods III

- **N-gram-based techniques**:
  - compute the similarity between two strings,
  - counting the number of similar n-grams they share,
  - the more similar n-grams between two strings exist the more similar they are.
- Similarity coefficient: $\delta_n(a, b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$
  - α and β are the n-gram sets for two words *a* and *b*,
  - $|\alpha \cap \beta|$ - similar n-grams, $|\alpha \cup \beta|$ - unique n-grams.

Example

fact vs. **fract**

Bigrams in fact : -f fa ac ct t- 5 bigrams
Bigrams in fract : -f fr ra ac ct t- 6 bigrams
$|\alpha \cap \beta|$: -f ac ct t- 4 bigrams
$|\alpha \cup \beta|$: -f fa fr ra ac ct t- 7 bigrams

$$\delta_2(fact, fract) = \frac{4}{7} = 0.57$$

Isolated-word methods IV

- **Rule-based techniques:**
  - a set of rules that capture common spelling and typographic errors,
  - "inverses" of common errors,
  - applying these rules to the misspelled word.
- **Probabilistic techniques:**
  - based on statistical features of the language,
  - estimated by collecting n-gram frequency statistic on a large corpus,
    - *transition probabilities* - probability that a given letter will be followed by another given letter,
    - *confusion probabilities* - how often a given letter is mistaken or substituted for another given letter.
- **Neural networks:**
  - based on back-propagation networks,
  - input node for every possible n-gram in every position of the word,
  - output node for each word in the dictionary.
  - new and promising technique in spell-checking

Grammar checking

Example

"That's good to now"
"That's good to know"

- Grammar checking starts where spell checking ends,
- deals with the most difficult and complex type of language errors,
  - wrong word order,
  - verb tense errors,
  - subject/verb agreement,
  - punctuation errors,
  - etc...
- Two main approaches:
  - *rule-based methods* - very time-consuming and with an inherent lack of flexibility.
  - *statistical methods* - easier and faster to implement, more flexible and adaptable [Nazar and Renau, 2012].

Rule-based grammar checking

- Testing the input text against a set of handcrafted rules.

Example

Rule: I + Verb (3rd person, singular form)
Corresponds to the **incorrect** verb form usage – "I has a dog"

- Advantages:
  - rules can be easily added, modified or removed,
  - rule can have a corresponding extensive explanation,
  - easily debuggable – decisions can be traced to a particular rule,
  - rules can be authored by the linguists, possessing limited or no programming skills.
- Disadvantages:
  - large amount of manual work,
  - needed to build an extensive rule set [Mozgovoy, 2011].

Rule-based grammar checker LanguageTool

How does a rule-based grammar checker **LanguageTool**[2] work [Naber, 2003]?

1. Takes plain text as input.
2. Splits text into sentences.
3. Splits sentences into words.
4. Finds part-of-speech tags for each word and its base form (walks - walk).
5. Matches the analyzed sentences against error patterns and runs rules.

[2] `https://languagetool.org/`

Rule example in LanguageTool

Example

"I thing that's a good idea."

```
<rule id="YOU_THING" name="Possible typo 'I/you/... thing(think)'">

<pattern mark_from="1">
        <token regexp="yes">I|you</token>
        <token regexp="yes">thing|things</token>
</pattern>

<message>Did you mean <suggestion>think</suggestion> ?</message>
<example type="correct">I <marker>think</marker> that's a good idea.</example>

</rule>
```

Statistical grammar checking

- Based on analysis of grammatically correct POS-annotated corpus,
- build a list of POS tag sequences,
  - some sequences are very common (determiner, adjective, noun as in *the old man*)
  - others will probably not occur at all (determiner, determiner, adjective)
- sequences which occur often in the corpus are considered correct,
- uncommon sequences might be errors.

Word completion

- Reduce the number of keystrokes,
- suggesting the completion of the word,
- use context information to predict what block of characters (letters, n-grams, syllables, words, or entire phrases) a person is going to write next,
- two goals:
  - store a wide-coverage word or language model,
  - prediction at earliest possible point of a character sequence being entered [Van den Bosch, 2011].

Best results

- Spell checking (first suggestion):
  - English – 95 % [Brill and Moore, 2000],
  - Czech – 73 % [Richter et al., 2012],
- Grammar checking (various tests average):
  - English – 55 % [Nazar and Renau, 2012],
  - Czech – 40% [Petkevič, 2014],

References I

Brill, E. and Moore, R. C. (2000).
An improved error model for noisy channel spelling correction.
In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gupta, N. and Mathur, P. (2012).
Spell checking techniques in nlp: A survey.
International Journal of Advanced Research in Computer Science and Software Engineering, 2(12).

Hladek, D., Stas, J., and Juhar, J. (2013).
Unsupervised spelling correction for slovak.
Advances in Electrical and Electronic Engineering, 11(5):392–397.

References II

Holmes, D. and McCabe, M. C. (2002).
Improving precision and recall for soundex retrieval.
In *Information Technology: Coding and Computing, 2002. Proceedings. International Conference on*, pages 22–26. IEEE.

Mozgovoy, M. (2011).
Dependency-based rules for grammar checking with languagetool.
In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 209–212.

Naber, D. (2003).
A rule-based style and grammar checker.

References III

Nazar, R. and Renau, I. (2012).
Google books n-gram corpus used as a grammar checker.
In *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*, EACL 2012, pages 27–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petkevič, V. (2014).
Kontrola české gramatiky (český grammar checker).
Studie z aplikované lingvistiky - Studies in Applied Linguistics, 5(2):48–66.

Richter, M., Straňák, P., and Rosen, A. (2012).
Korektor-a system for contextual spell-checking and diacritics completion.
In *COLING (Posters)*, pages 1019–1028.

References IV

Van den Bosch, A. (2011).
Effects of context and recency in scaled word completion.
Computational Linguistics in the Netherlands Journal, 1.

Whitelaw, C., Hutchinson, B., Chung, G. Y., and Ellis, G. (2009).
Using the web for language independent spellchecking and autocorrection.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 890–899, Stroudsburg, PA, USA. Association for Computational Linguistics.