RÓBERT HORVÁTH
Supervisor: MARIÁN ŠIMKO
STU FIIT

SUPPORTING TERM EXPLANATION WHILE BROWSING IN SLOVAK

NECESSITY of full understanding of web page

NEED to choose the proper meaning by ourselves

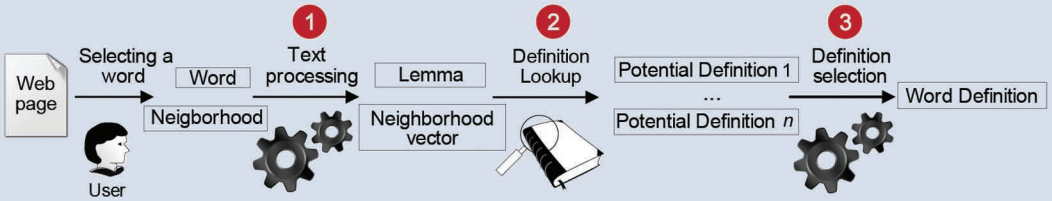
MISSING tool for neighborhood based explanation

WORD NEIGHBORHOOD · LEMMA · COSINE SIMILARITY · TEXT EXPLANATION · DEFINITION LOOKUP · WORD DISAMBIGUATION · TEXT PROCESSING

SOLUTION

METHOD FOR DISAMBIGUATED TERM DEFINITION ACQUISITION

1. web page text pre-processing
2. potential definitions lookup
3. correct definition selection



Web page pre-processing

- HTML tags removal
- Lemmatization - transforming a word into its basic form. We use Lematizér JUĽŠ SAV with 96.1% accuracy.
- Stop words removal - while it does not affect text meaning.
- Text segmentation and representation Bag of Words model.

Potential definition lookup

- Online dictionaries
- Lemma as an input
- Potential definitions output (polysemy)

Correct definition selection

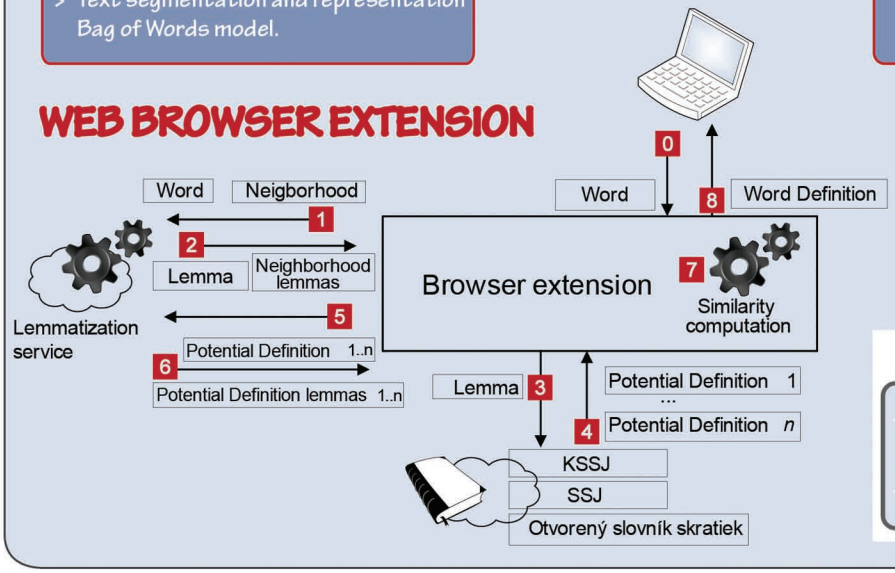
Word neighborhood options:

- whole text
- paragraph only
- sentence only

Similarity calculation:

$$CosSim(D_A, D_B) = \cos(\theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}.\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

WEB BROWSER EXTENSION



1. word lemma and link to dictionary
2. meaning index and meaning switcher
3. word meaning and samples of usage
4. feedback
5. synonyms

PRELIMINARY EXPERIMENT

- different approaches to word neighborhood selection
- experiment to find the best

	Correct word definition selection accuracy
Whole text	70%
Paragraph only	80%
Sentence only	60%

CORRECT TERM DEFINITION ACQUSITION

- live experiment with 20+ users
- gathering explicit feedback
- 1200 logs

Single meaning words	95,34%
Multiple meaning words	68,67%
Other than first meaning	63,85%
Overall	77,11%