Automatic Construction of WordNets by Using Machine Translation and Language Modeling

Martin Saveski, Igor Trajkovski

Information Society
Language Technologies
Ljubljana 2010

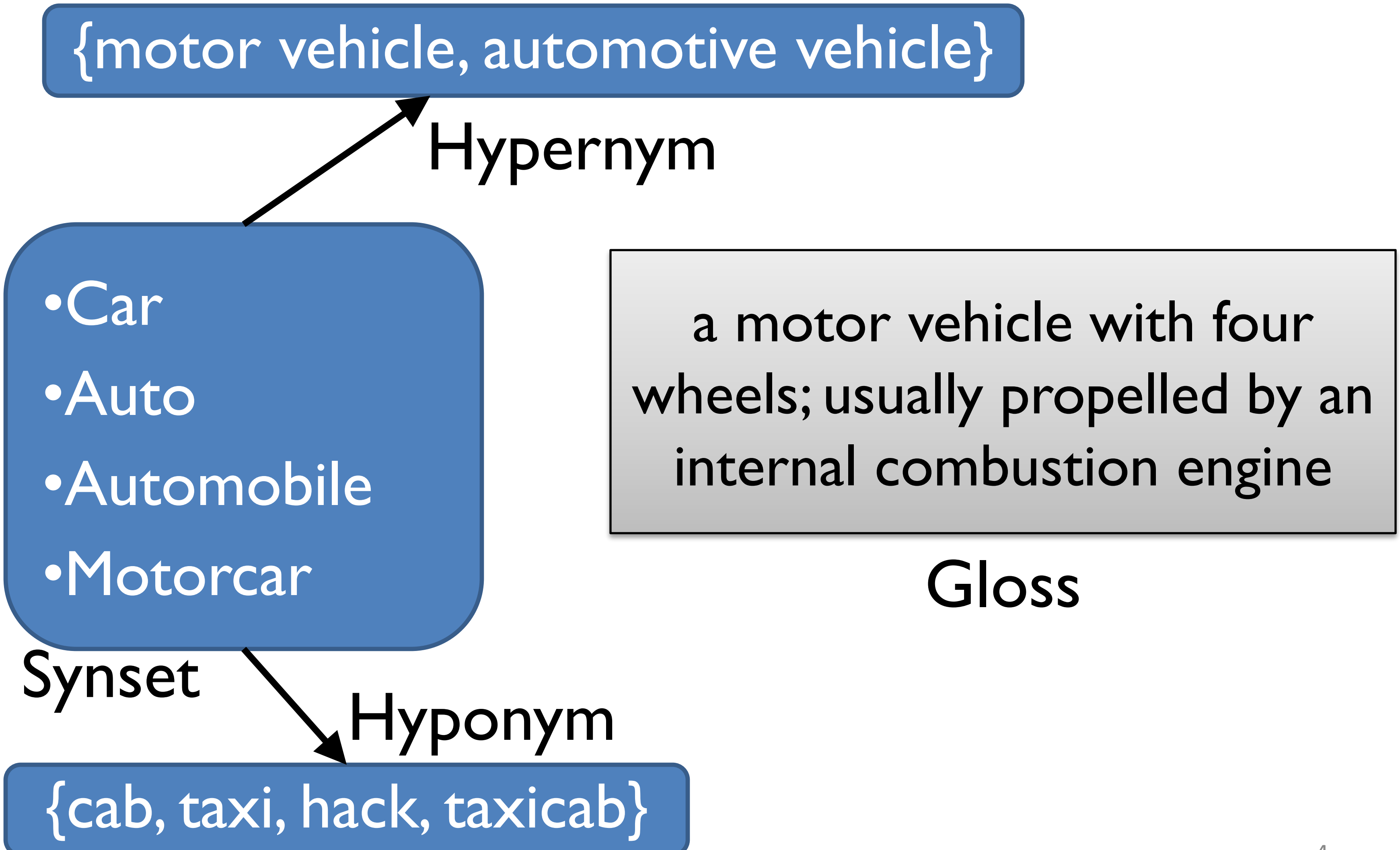
Outline

- WordNet
- Motivation and Problem Statement
- Methodology
- Results
- Evaluation
- Conclusion and Future Work

WordNet

- Lexical database of the English language
- Groups words into sets of cognitive synonyms called *synsets*
- Each synsets contains *gloss* and *links* to other synsets
  - Links define the place of the synset in the conceptual space
- Source of motivation for researchers from various fields

WordNet Example

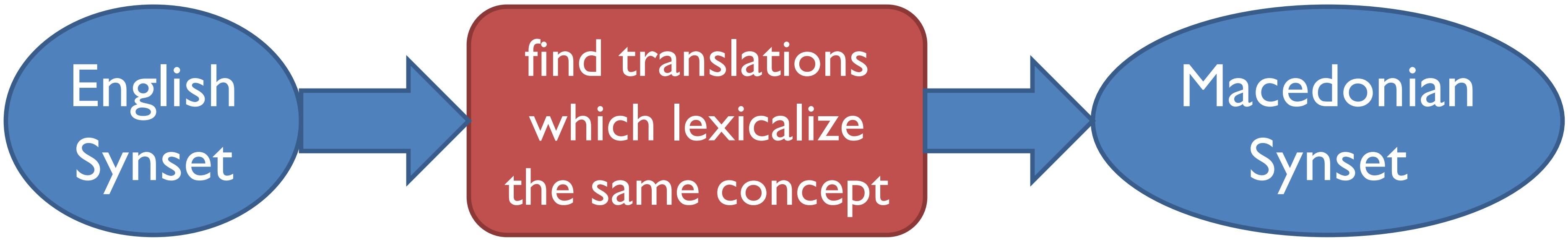


Motivation

- Plethora of WordNet *applications*
  - Text classification, clustering, query expansion, etc.
- There is no publicly available WordNet for the Macedonian Language
  - Macedonian was not included in the EuroWordNet and BalkaNet projects
- Manual construction is expensive and labor intensive process
  - Need to automate the process

Problem Statement

- Assumptions:
  - The conceptual space modeled by the PWN is not depended on the language in which it is expressed
  - Majority of the concepts exist in both languages, English and Macedonian, but have different notations



Given a synset in English, it is our goal to find a set of words which lexicalize the same concept in Macedonian

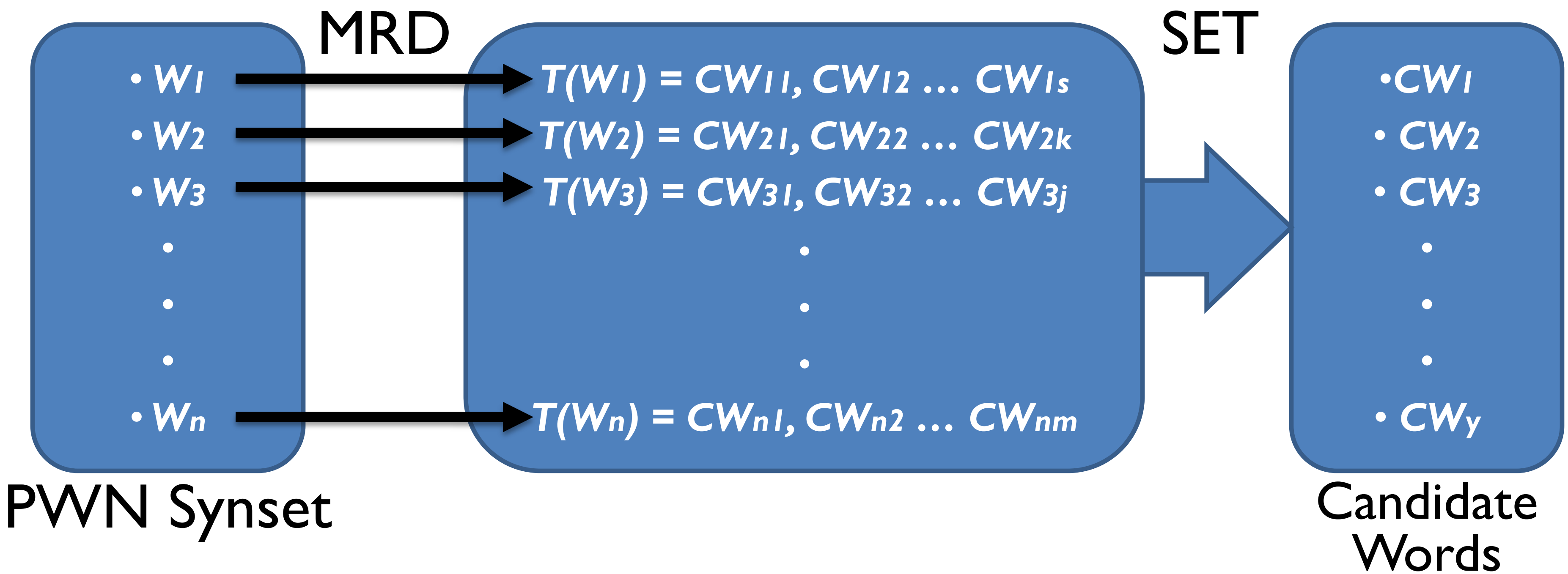
Resources and Tools

- Resources:
  – Princeton implementation of WordNet (PWN) – *backbone* for the construction
  – English-Macedonian Machine Readable Dictionary (in-house-developed) – 182,000 entries
- Tools:
  – Google Translation System (Google Translate)
  – Google Search Engine

Methodology

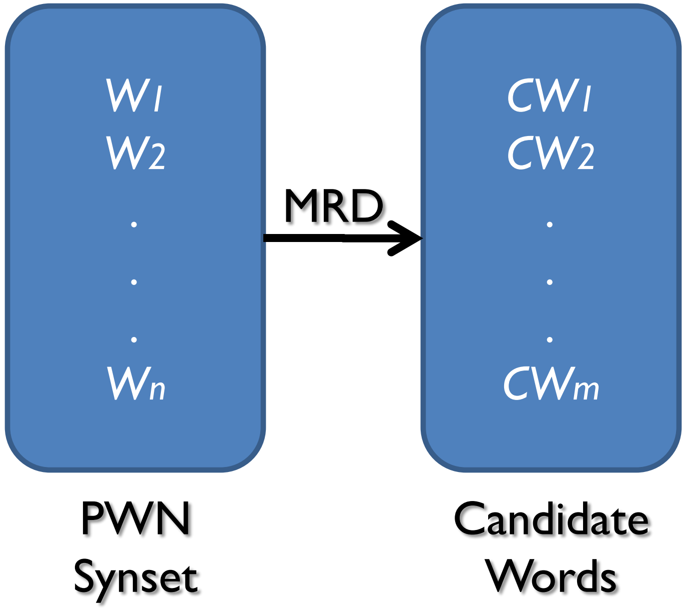
1. Finding Candidate Words
2. Translating the synset gloss
3. Assigning scores the candidate words
4. Selection of the candidate words

Finding Candidate Words



- $T(W_1)$ contains translations of all senses of the word W_1
- Essentially, we have Word Sense Disambiguation (WSD) problem

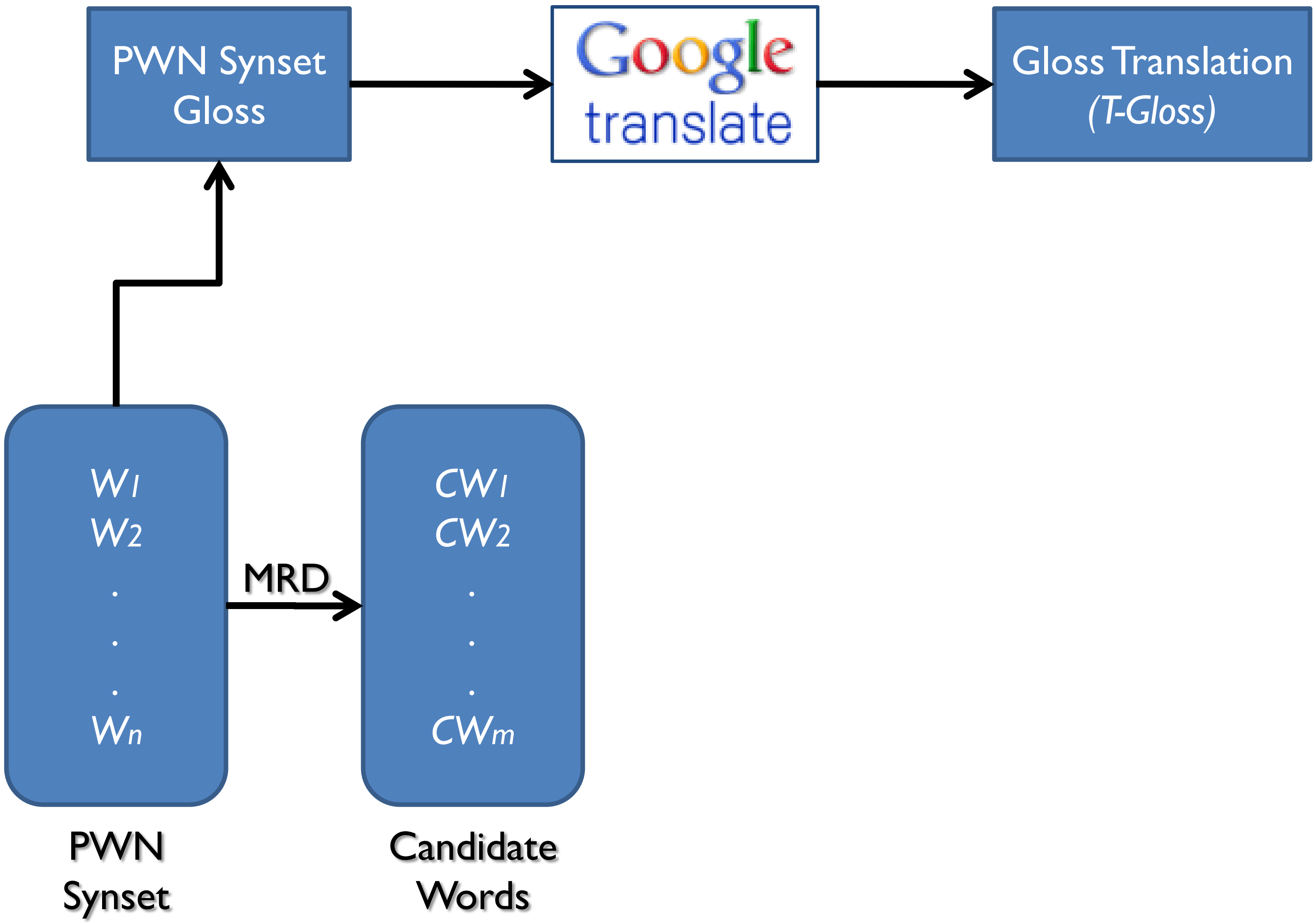
Finding Candidate Words (cont.)



Translating the synset gloss

- Statistical approach to WSD:
  - Using the word sense definitions and a large text corpus, we can determine the sense in which the word is
- Word Sense Definition = Synset Gloss
- The gloss translation can be used to measure the *correlation* between the synset and the candidate words
- We use *Google Translate* (EN-MK) to translate the glosses of the PWN synsets

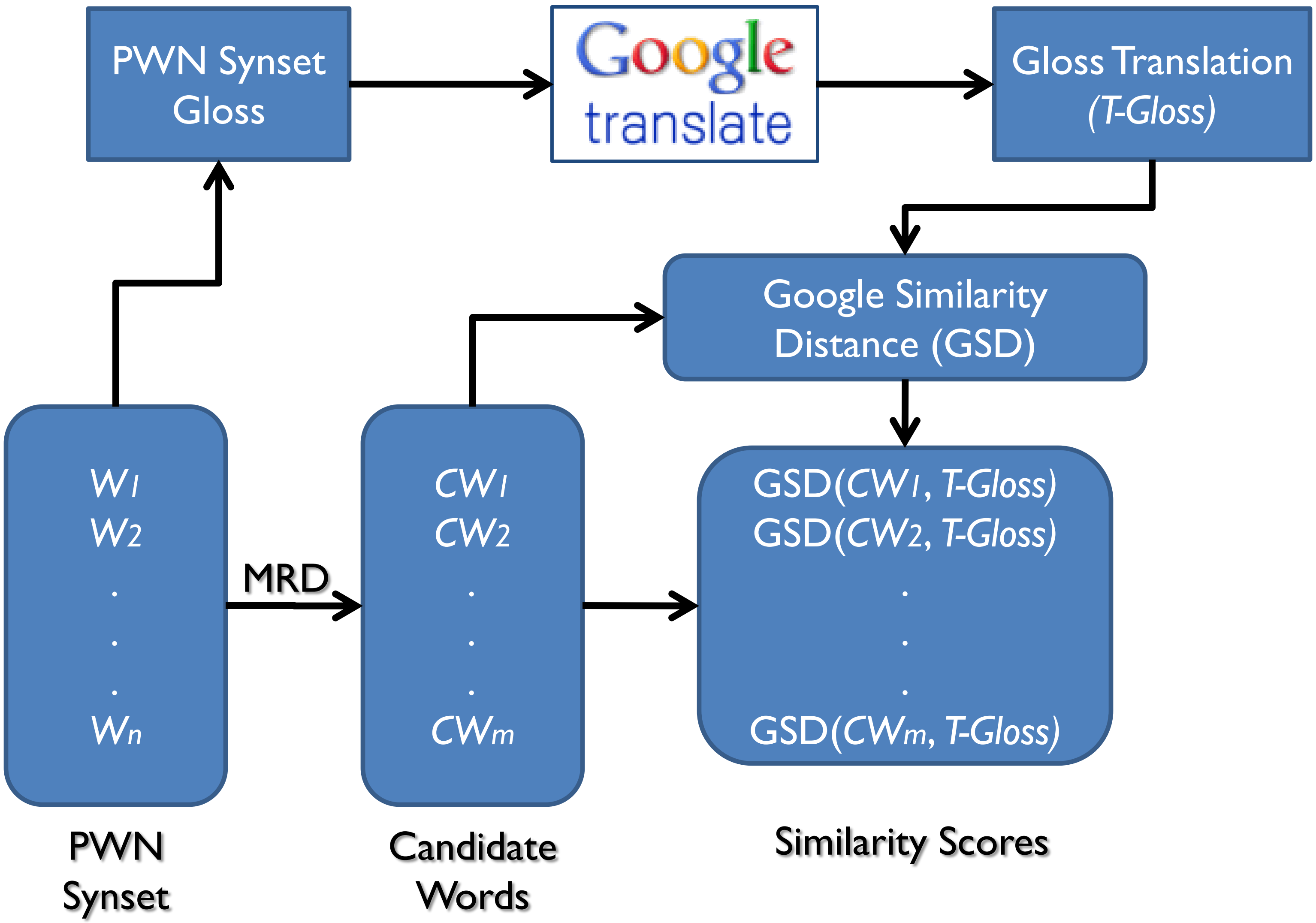
Translating the synset gloss (cont.)



Assigning scores to the candidate words

- To apply the statistical WSD technique we lack a large, domain independent text corpus
- Google Similarity Distance (GSD)
  - Calculates the semantic similarity between words/phrases based on the Google result counts
- We calculate GSD between each candidate word and gloss translation
- The GSD score is assigned to each candidate word

Assigning scores to the candidate words



Selection of the candidate words

- Selection by using two thresholds:

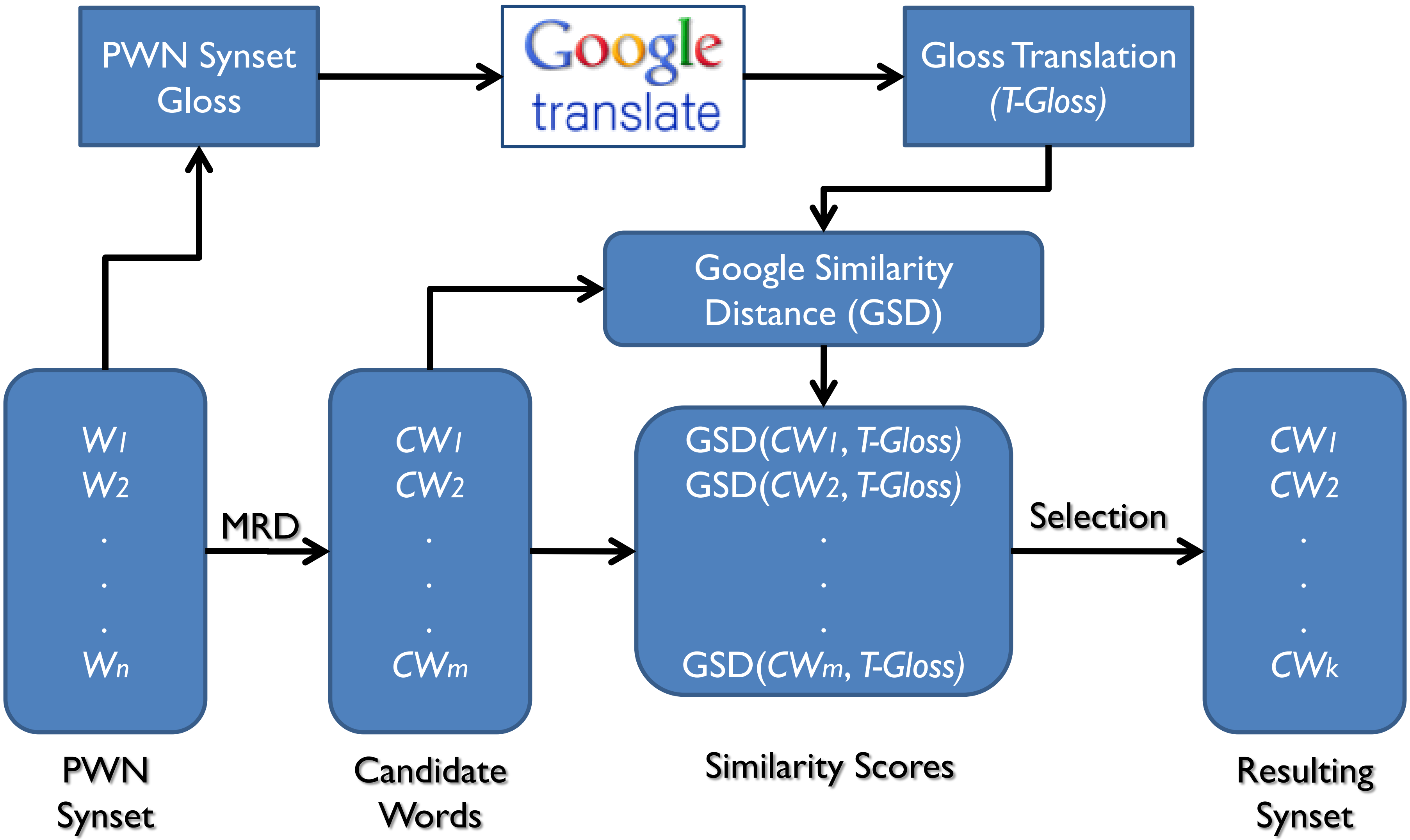
  1. *Score(CW) > T_1*

     - Ensures that the candidate word has minimum correlation with the gloss translation

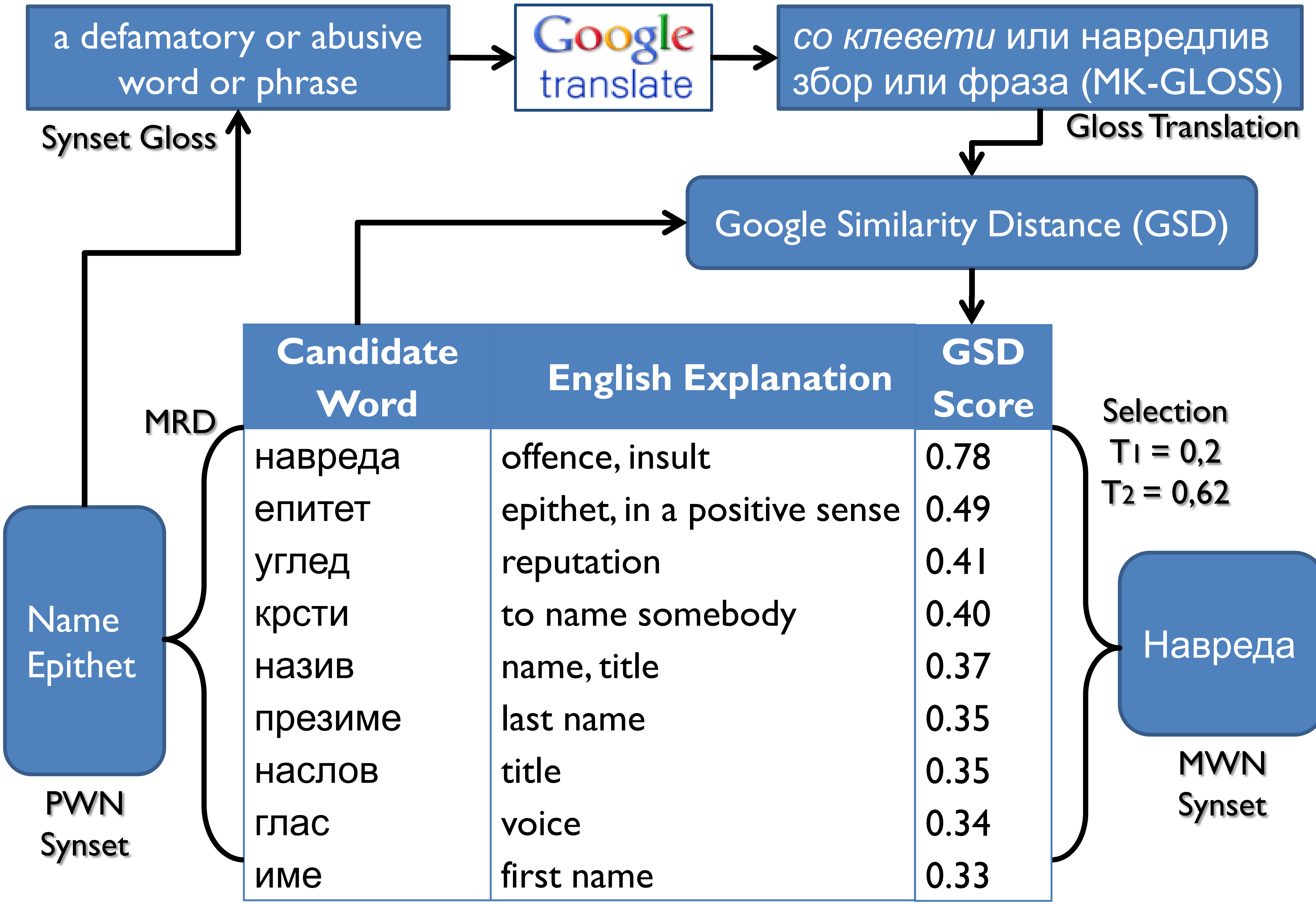
  2. *Score(CW) > (T_2 x MaxScore)*

     - Discriminates between the words which capture the meaning of the synset and those that do not

Selection of the candidate words (cont.)

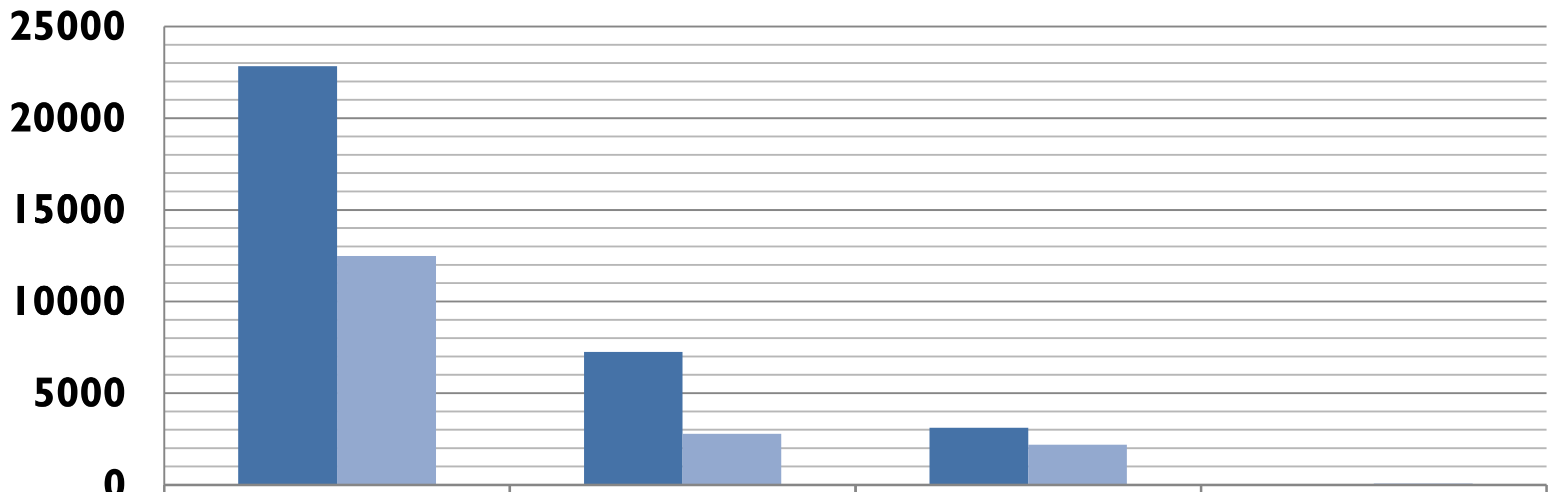


Example



Candidate Word	English Explanation	GSD Score
навреда	offence, insult	0.78
епитет	epithet, in a positive sense	0.49
углед	reputation	0.41
крсти	to name somebody	0.40
назив	name, title	0.37
презиме	last name	0.35
наслов	title	0.35
глас	voice	0.34
име	first name	0.33

Results of the MWN construction



	Nouns	Verbs	Adjectives	Adverbs
Synsets	22838	7256	3125	57
Words	12480	2786	2203	84

Size of the MWN

NB: All words included in the MWN are lemmas

Evaluation of the MWN

- There is no manually constructed WordNet (lack of Golden Standard)
- Manual evaluation:
  - Labor intensive and *expensive*
- Alternative Method:
  - Evaluation by use of MWN in practical applications
  - MWN applications were our motivation and ultimate goal

MWN for Text Classification

- Easy to measure and compare the performance of the classification algorithms
- We extended the synset similarity measures to word-to-word i.e. text-to-text level
  – *Leacock and Chodorow* (LCH) (node-based)
  – *Wu and Palmer* (WUP) (arc-based)
- Baseline:
  – *Cosine Similarity* (classical approach)

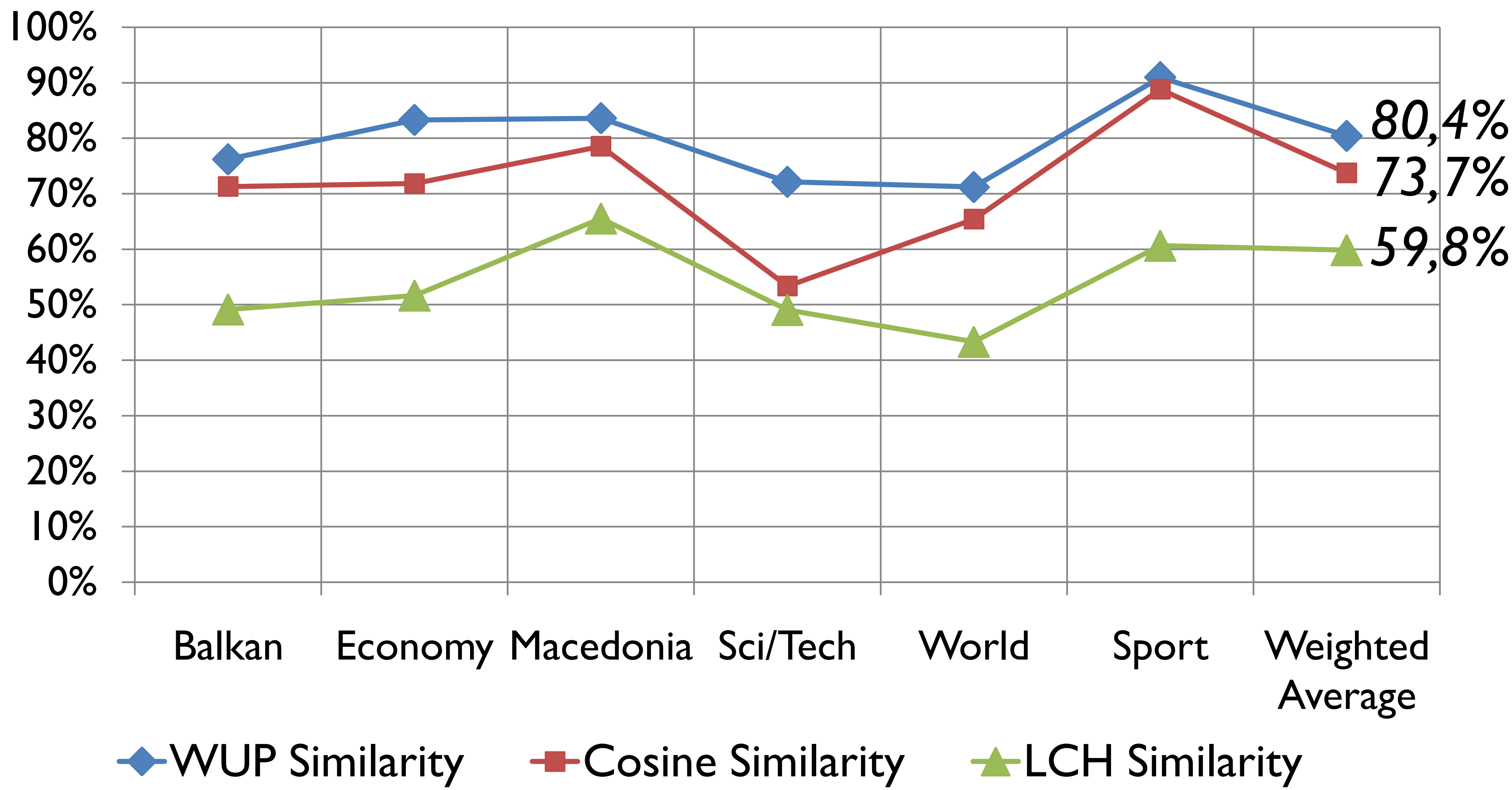
MWN for Text Classification (cont.)

- Classification Algorithm:
  - K Nearest Neighbors (*KNN*)
  - Allows the similarity measures to be compared unambiguously
- Corpus: A1 TV - News Archive (2005-2008)

Category	Balkan	Economy	Macedonia	Sci/Tech	World	Sport	TOTAL
Articles	1,264	1,053	3,323	920	1,845	1,232	9,637
Tokens	159,956	160,579	585,368	17,775	222,560	142,958	1,289,196

A1 Corpus, size and categories

MWN for Text Classification – Results



Text Classification Results (F-Measure, 10-fold cross-validation)

Future Work

- Investigation of the semantic relatedness between the candidate words
  - Word Clustering prior to assigning to synset
  - Assigning group of candidate words to the synset
- Experiments of using the MWN for other applications
  - Text Clustering
  - Word Sense Disambiguation

Q&A

Thank you for your attention.

Questions?

Google Similarity Distance

- Word/phrases acquire meaning from the way they are used in the society and from their *relative semantics* to other words/phrases
- Formula:

$$GSD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

f(x), f(y), f(x,y) – results counts of x, y, and (x, y)

N – Normalization factor

Synset similarity metrics

- Leacock and Chodorow (LCH)

$$sim_{LCH}(s_1, s_2) = -\log \frac{len\,(s_1, s_2)}{2 * D}$$

len – number of nodes form *s1* to *s2*,

D – maximum depth of the hierarchy

- Measures in number of nodes

Synset similarity metrics (cont.)

- Wu and Palmer (WUP)

$$sim_{WUP}(s_1, s_2) = \frac{2 * depth(LCS)}{depth(s_1) + depth(s_2)}$$

LCS – most specific synset ancestor to both synsets

- Measures in number of links

Semantic Word Similarity

- The similarity of W_1 and W_2 is defined as:
- The maximum similarity (minimum distance) between the:
  - Set of synsets containing W_1,
  - Set of synsets containing W_2

Semantic Text Similarity

- The similarity between texts T_1 and T_2 is:

$$sim(T_1\,, T_2) = \frac{1}{2}\left(\frac{\sum_{w\,\in\{T_1\}}\left(maxSim(w, T_2) * idf(w)\right)}{\sum_{w\,\in\{T_1\}} idf(w)} + \frac{\sum_{w\,\in\{T_2\}}\left(maxSim(w, T_1) * idf(w)\right)}{\sum_{w\,\in\{T_2\}} idf(w)}\right)$$

  - *idf* – inverse document frequency (measures word specificity)