Fighting arbitrariness in WordNet-like lexical databases – A natural language motivated remedy

Shun Ha Sylvia Wong

Computer Science

Aston University

U.K.

s.h.s.wong@aston.ac.uk

Lexical Databases

Lexical Database: a database of words $\approx$ a dictionary (of terms)

- HowNet (`http://www.keenage.com/html/e_index.html`)
- WordNet (`http://www.cogsci.princeton.edu/~wn/`)
- EuroWordNet (EWN) (`http://www.illc.uva.nl/EuroWordNet/`)
- Chinese Concept Dictionary (CCD)

  (`icl.pku.edu.cn/yujs/papers/pdf/intr2CCD.pdf`)

Differ In:

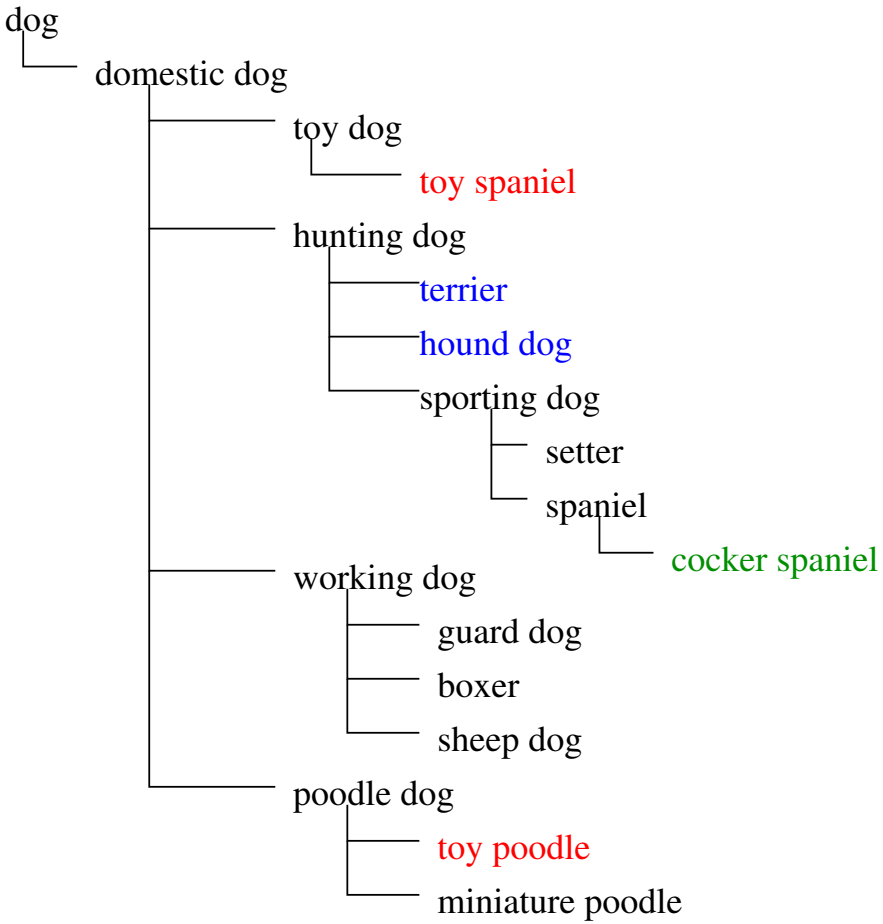
- The detailed organizations of real-world concepts,
- the set of concept relations, and
- how the knowledge base is structured.

In Common:

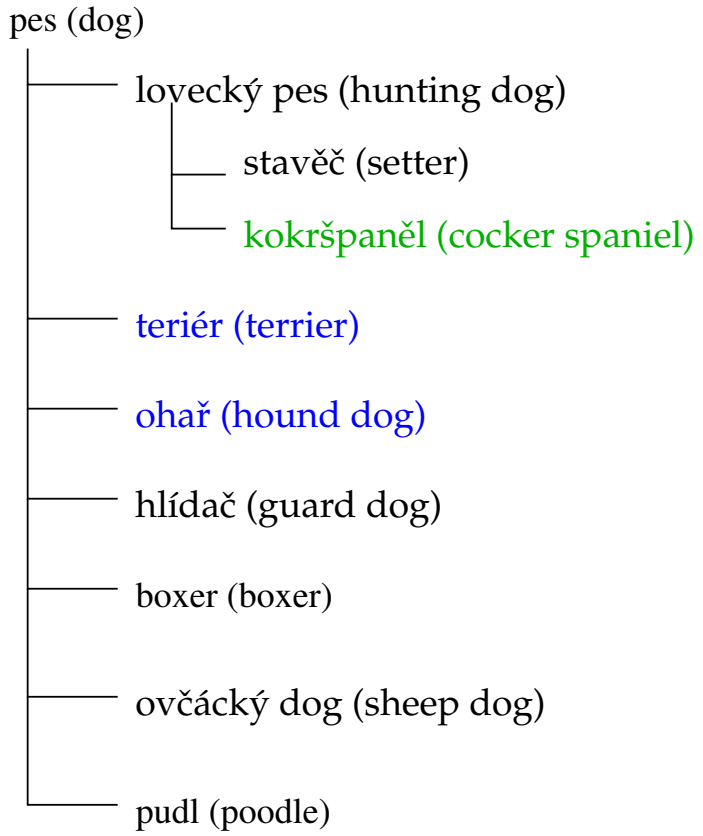
- Contains an ontology of [supposedly language-independent] concepts.
- Relates concepts by a set of predefined relations.
- Classification of concepts is, by and large, manual-driven.

Concept classification – an arbitrary process?

WordNet 1.5



EuroWordNet 2: Czech WordNet



Subjective association of concepts and relations leads to arbitrariness which results in a fragmented and incoherent knowledge base.

1. *Incoherent concept classification*

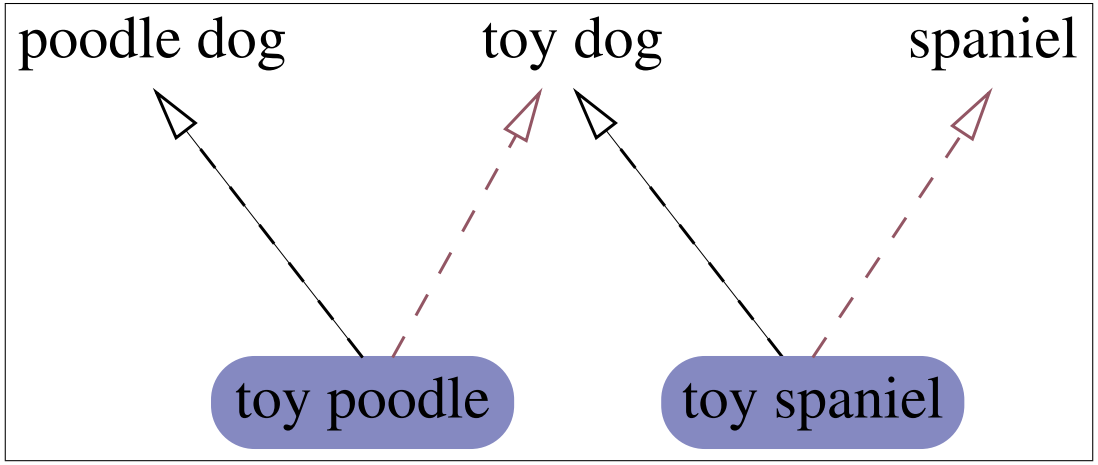


Figure 1: `toy dog`, `poodle dog` or `spaniel`?

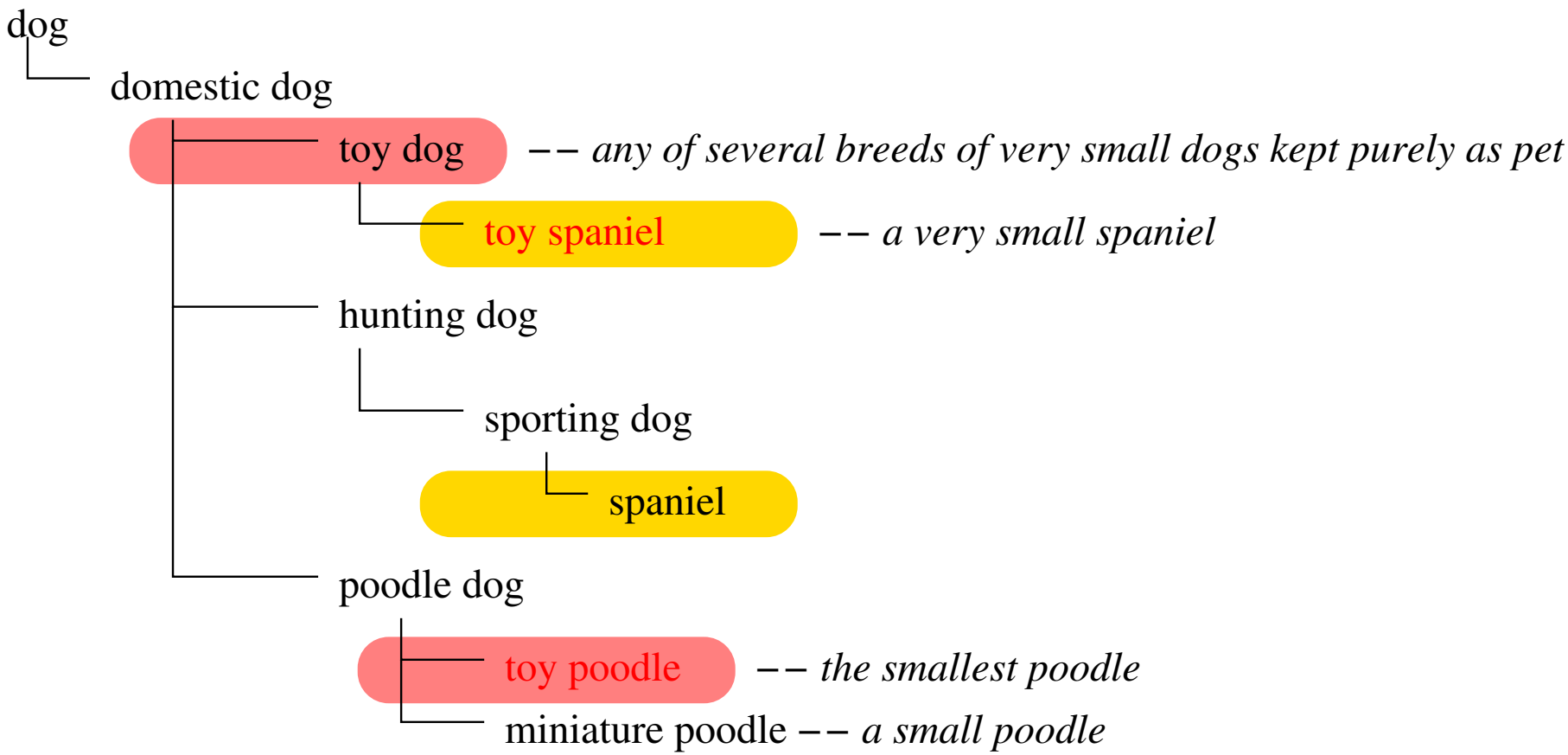


Figure 2: Classification of `toy dog` and `poodle dog` in WordNet 1.5

2. *Mis-classified concepts*

Within EuroWordNet:

`lovecký pes` (*hunting dog*)	$\equiv$	`sporting dog`

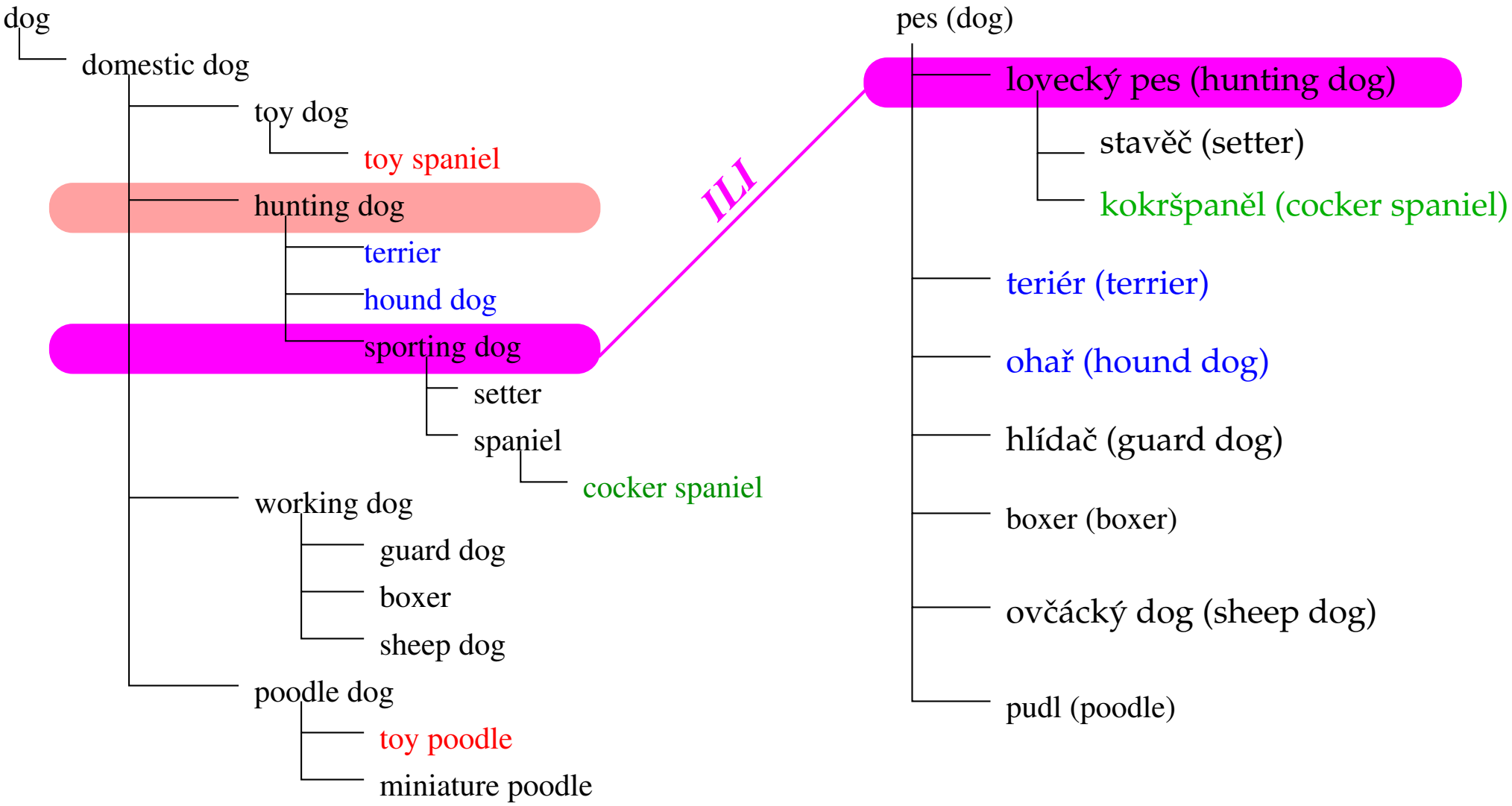


Figure 3: EuroWordNet 2: `lovecký pes` (*hunting dog*) ≡ `sporting dog`

A Natural Language Motivated Remedy

- Natural languages facilitate a concise communication of real world concepts using sequences of symbols.

  ☞ *Could knowledge representation in natural languages be employed to alleviate arbitrariness in lexical databases?*

- Words are formed by concatenating morphemes.
- Chinese words often display sufficient word formation information that meaningful grouping of Chinese words can easily be formed using their component characters. (Cf. Figure 4)

  ☞ *Morphological structure of words provide clues to concept organization.*

Spotting the relations…

What does each column of words have in common?

頭髮 (*hair*)	牙膏 (*toothpaste*)	戰車 (*chariot*)
假髮 (*wig*)	牙刷 (*toothbrush*)	戰士 (*warrior*)
長假髮 (*peruke*)	牙線 (*dental floss*)	戰利品 (*plunder*)
長髮 (*long hair*)	牙醫 (*dentist*)	戰役 (*military campaign*)
短髮 (*short hair*)		戰鬥 (*fight*)
直髮 (*straight hair*)		戰鬥者 (*fighter*)
曲髮 (*wavy hair*)		戰鬥機 (*fighter [plane]*)
鬈髮 (*curly hair*)		噴射式戰鬥機 (*fighter jet*)

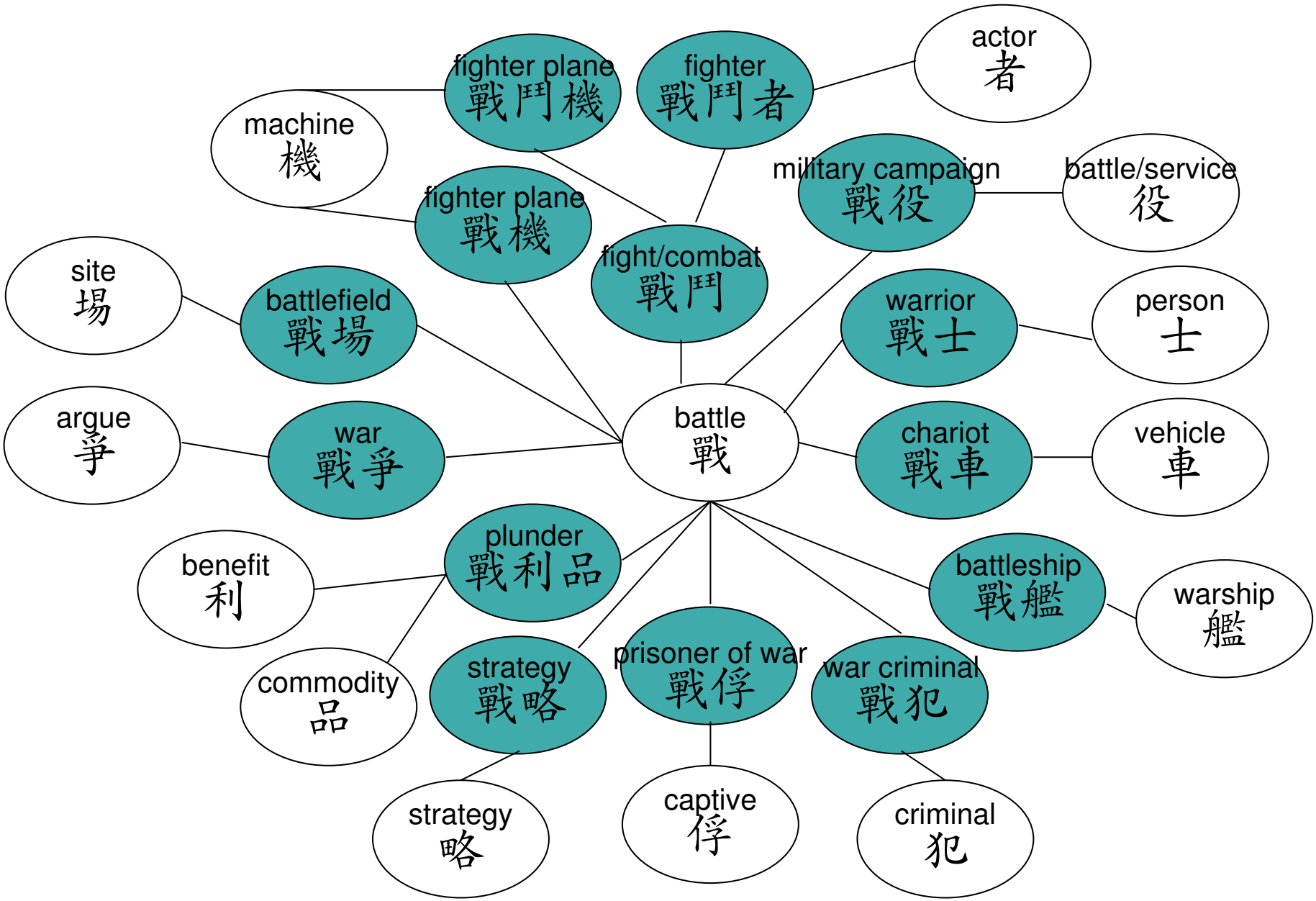


Figure 4: 戰 (*battle/war*) and some battle-related Chinese words

Exploiting concept relatedness in Chinese for WSD

- Chinese words are arranged naturally in clusters. (Cf. Figure 4)
- Each concept cluster reveals the context in which all member concepts exist. (Cf. Figure 4)
- Concept relatedness in Chinese enables Context-based Word Sense Disambiguation (WSD). (Cf. Figure 5)

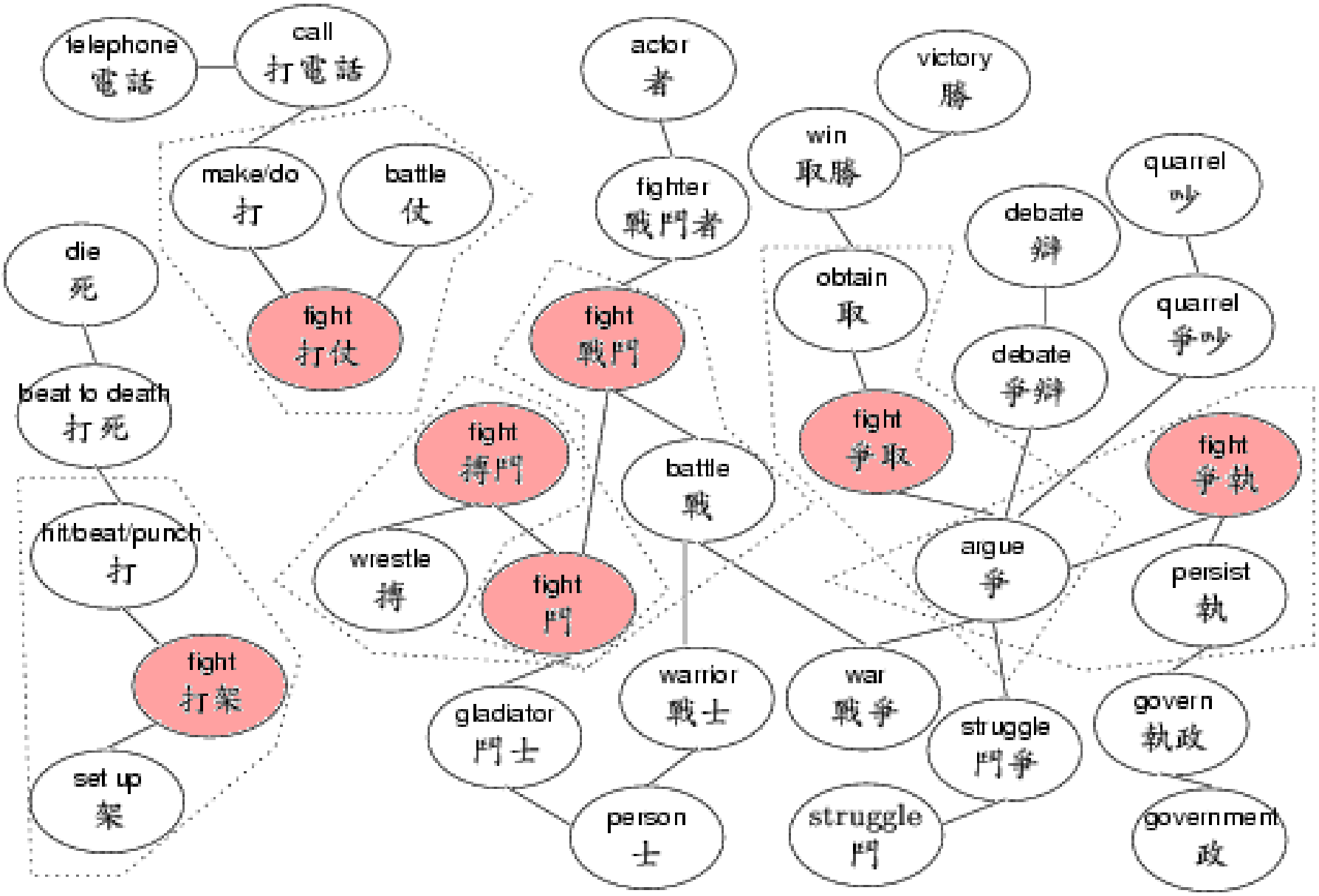


Figure 5: Various senses of 'fight' and their Chinese counterparts

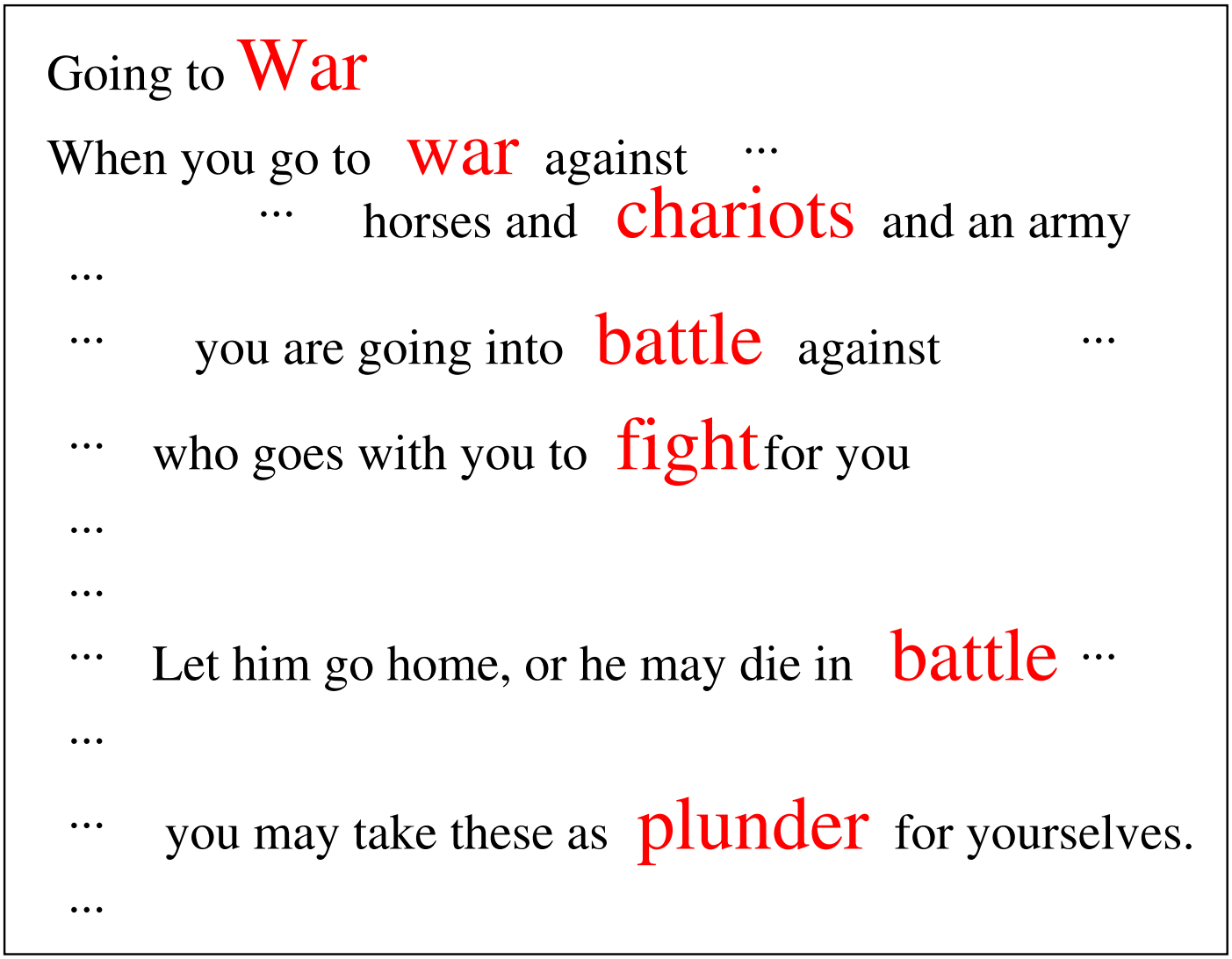


Figure 6: 'fight' in battle/war context (Deuteronomy 20:1-20)

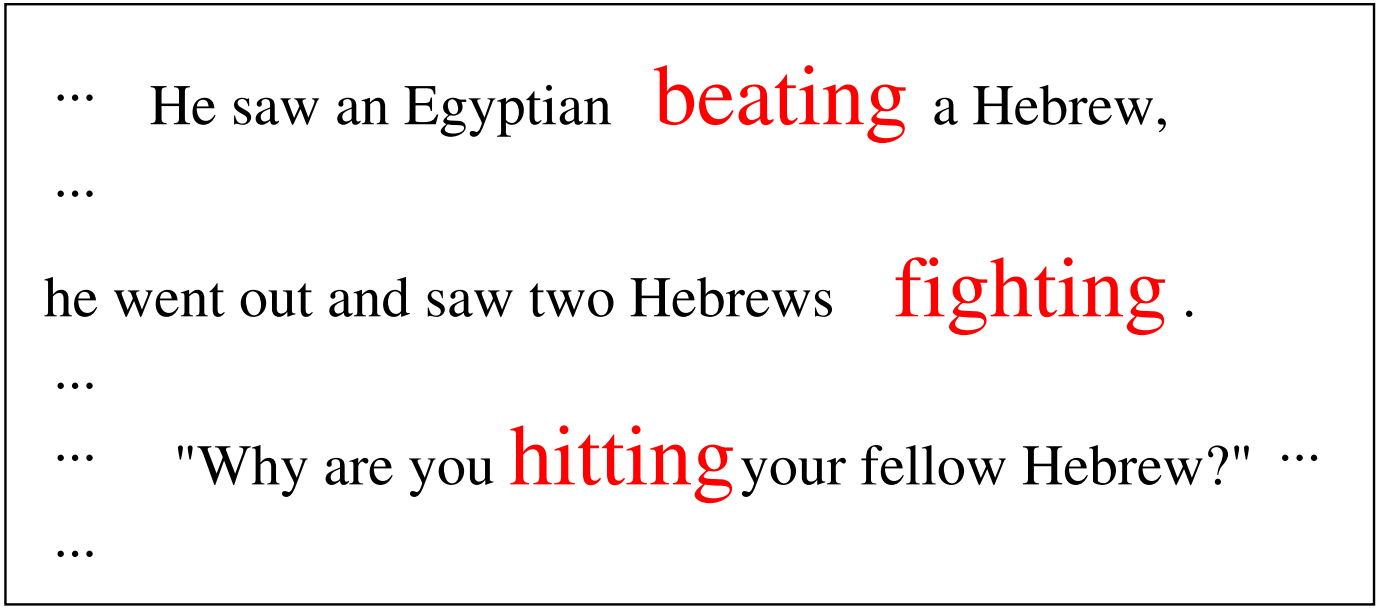


Figure 7: 'fight' in hitting/punching context (Exodus 2:11-14)

Require

- Chinese-English dictionary

  currently 2566 Chinese-English word pairs

- Lemmatiser

  TreeTagger from IMS Stuttgart (Schmid 1996)

- English text

  seven short extracts from the NIV Bible (International Bible Society 1983)

- Java

Procedure

- Text Preprocessing
  - lemmatized input text
- Dictionary Lookup
  - based on lexical units of < 5 English words
  - locate all Chinese counterparts
- Word Sense Selection
  - locate five most frequently occurred characters
  - for each lexical unit, select the Chinese counterpart(s) which comprises a frequently occurred character

Results

- 45 lexical units disambiguated, 37 of them correctly interpreted and 3 of them didn't contain the best available interpretations.
- Correctly disambiguated lexical units: fight, beat, blow, chariot, hit, loss, march, officer, plunder, strike
- 87.5% correctness

  (*based on simple word counting and even without considering POS info.*)

Conclusion

- Existing lexical databases are vulnerable to arbitrariness.
- Concept formation in natural languages has potential to combat arbitrariness in lexical databases.
- Concept relatedness in Chinese can be exploited to perform Word Sense Disambiguation.

arbitrary Based on or subject to individual judgment or preference (Houghton Mifflin Company 2000)

Ontology `http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=ontology`

1. philosophy A systematic account of Existence.
2. artificial intelligence (From philosophy) An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them.
3. information science The hierarchical structuring of knowledge about things by subcategorising them according to their essential (or at least relevant and/or cognitive) qualities.

References

Houghton Mifflin Company (2000), 'The american heritage dictionary of the english language', [Online]. Available at: http://www.yourdictionary.com/ [14 May , 2003].

International Bible Society, ed. (1983), *The NIV (New International Version) Bible*, 2nd edn, Zondervan, Grand Rapids, MI. [Online]. Available at: http://bible.gospelcom.net/ [13 August, 2003].

Schmid, H. (1996), 'Treetagger – a language independent part-of-speech tagger', [Online]. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html [13 August, 2003]. A tool developed within the project on Textual Corpora and Tools for their Exploration at IMS Stuttgart.