Towards a Hybrid Approach to Word Sense Disambiguation in Machine Translation

Márton Miháltz

MorphoLogic

Orbánhegyi út 5.

Budapest, Hungary

Introduction

- Word Sense Disambiguation:

  Select the adequate ***sense*** *of a polysemous lexical item in its context, from a pre-specified sense inventory.*

- **WSD** in Machine Translation:

  Select the correct ***translation*** *in the target language for an ambiguous item in the source language, based on its context in the translation unit (TU).*

- A Hungarian-English **example**:

a) *A* ***nap*** *melegen sütött.*	The **SUN** was shining.
b) *Három* ***nap*** *telt el.*	Three **DAYs** passed.

- WSD doesn't (shouldn't) deal with POS ambiguities

Introduction 2.

- But **why** use WSD in MT?

  Needed in **rule-based** approach (vs. **statistical** MT)

- MorphoLogic's **MetaMorpho** English-Hungarian MT system
  - **Translation patterns** (constructions): manually created analysis & generation CF **rule-pairs** (now over 100,000 rules)
  - **Some verbs: several** translations, disambiguated by grammar

They **fired** the furniture[-ANIM].	***Eltüzelték** a bútort.*
He **fired** the employee[+ANIM].	***Kirúgta** az alkalmazottat.*

  - **Most** cases (most verbs, all **nouns**): **Single** translation: the most frequent SL sense. Problem with **polysemous** lexical items!

We moved to another **state**.	*Egy másik **államba** költöztünk.*
Her **state** was satisfactory.	*Az **állapota / *állama** kielégítő volt.*

  - Collocations with polysemous words have their own rules:

„**state** of affairs"	***helyzet***

Our approach

- Disambiguate, i.e. select correct translation using **context** of ambiguous word in SL translation unit
- **Supervised ML**: learn from sense-tagged SL examples
- Train **classifiers** for each polysemous word
- A **Hybrid** MT system:
  - **Rules** deal with **unambiguous** lexical units, multiword phrases
  - Some ambiguities resolved by syntax (rules)
  - **Statistical** WSD module disambiguates **polysemous** lexical items
  - Also manual word-sense disambiguation rules
  - **Translation Memory** (Hodász & Pohl, this vol.)
- **WSD module** specifies value of a grammar feature in SL analysis phase that will select correct translation in TL generation phase.

Contextual Features

- (Leacock et al, 1998)
- **Local** information:
  - Surface form of ambiguous word

    arm vs. ***arms***
  - Function words in 2+2 window

    behind the **church**
  - Open-class words in 3+3 window

    the **party** *won* the *elections*
- **Topical** information:
  - Bag of open-class words in whole context

    … *airport* … **plane** …
- Context words are **lemmatized**
- Feature values form **feature vectors**
- Different feature subsets possible for different polysemous items

The Classifier

- **Naïve Bayes** statistical ML algorithm
  - Simple to implement, fast & efficient
  - Performs well in NLP tasks, including WSD

    SensEval-3 English lexical sample task #1: HTSA3 system
  - Proved best in preliminary investigation
    - Compared to other **statistical** and **memory-based** learning schemes in **WEKA** toolkit
    - Precision with **10-fold cross-validation** on 1 dataset (OMWE *party.n*)
    - Naïve Bayes had highest precision with with **current feature configuration**. Other learning methods may work better with other feature combinations, or a different representation method.
    - No feature engineering (yet)

Dataset used in Experiment

- **Training corpora**: examples manually sense-tagged with WordNet senses, with context
  - SensEval-2, Open Mind Expert 1.0, line.n corpora
  - Experiment dataset: 42 polysemous English nouns
- WN senses **mapped** to Hungarian translations
  - 4 items: all English senses = 1 Hungarian translation
  - 34 items: #English senses > Hungarian translations
  - 4 items: #English senses < Hungarian translations

  Average sense count: 3,97 (English) → 2,49 (Hungarian)
- Preprocessing:
  - Segmentation, tokenization, POS-tagging, lemmatization
  - Removing unambiguous collocations formed with the polysemous target words

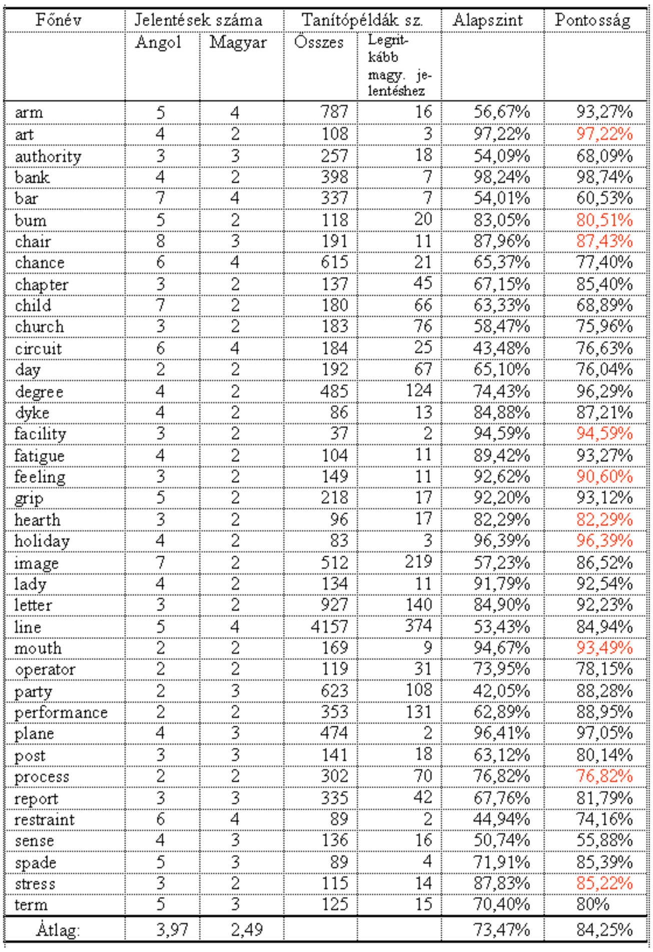
    Eg. "capital **letter**" Hungarian: always „***nagybetű***"

Evaluation of Experiment

- **10-fold cross-validation** on each of 38 nouns
- **Precision** =

  #(correctly tagged instances) / #(all instances)
- **Baseline** = majority sense
  - English: avg. **64.15%,** Hungarian avg.: **73.47%**
- **Average precision** (across 38 items):
  - With **English** sense tags: **76.39%**
  - With mapped **Hungarian** translations: **84.25%**
  - 9 items: precision <= baseline
- **Related** results:

  Leacock et al '98: Naïve Bayes, same features, *line.n*, 4.000 instances, **83%** precision, ours: **84,9%**



Főnév	Jelentések száma		Tanítópéldák sz.		Alapszint	Pontosság
	Angol	Magyar	Összes	Legrit-kább magy. je-lentéshez		
arm	5	4	787	16	56,67%	93,27%
art	4	2	108	3	97,22%	97,22%
authority	3	3	257	18	54,09%	68,09%
bank	4	2	398	7	98,24%	98,74%
bar	7	4	337	7	54,01%	60,53%
bum	5	2	118	20	83,05%	80,51%
chair	8	3	191	11	87,96%	87,43%
chance	6	4	615	21	65,37%	77,40%
chapter	3	2	137	45	67,15%	85,40%
child	7	2	180	66	63,33%	68,89%
church	3	2	183	76	58,47%	75,96%
circuit	6	4	184	25	43,48%	76,63%
day	2	2	192	67	65,10%	76,04%
degree	4	2	485	124	74,43%	96,29%
dyke	4	2	86	13	84,88%	87,21%
facility	3	2	37	2	94,59%	94,59%
fatigue	4	2	104	11	89,42%	93,27%
feeling	3	2	149	11	92,62%	90,60%
grip	5	2	218	17	92,20%	93,12%
hearth	3	2	96	17	82,29%	82,29%
holiday	4	2	83	3	96,39%	96,39%
image	7	2	512	219	57,23%	86,52%
lady	4	2	134	11	91,79%	92,54%
letter	3	2	927	140	84,90%	92,23%
line	5	4	4157	374	53,43%	84,94%
mouth	2	2	169	9	94,67%	93,49%
operator	2	2	119	31	73,95%	78,15%
party	2	3	623	108	42,05%	88,28%
performance	2	2	353	131	62,89%	88,95%
plane	4	3	474	2	96,41%	97,05%
post	3	3	141	18	63,12%	80,14%
process	2	2	302	70	76,82%	76,82%
report	3	3	335	42	67,76%	81,79%
restraint	6	4	89	2	44,94%	74,16%
sense	4	3	136	16	50,74%	55,88%
spade	5	3	89	4	71,91%	85,39%
stress	3	2	115	14	87,83%	85,22%
term	5	3	125	15	70,40%	80%
Átlag:	3,97	2,49			73,47%	84,25%

Discussion

- Experiment: in 9 of 38 cases, precision not exceeding or below baseline score
  - Variation in # of training instances – worst results:

    #(instances for least freq. sense) <= 20

    #(total instances) <= 200
  - Variation in context size (1-9 sentences)
  - Variation in context genre, style, ellaboration (newswire, AI assertions, web user input etc.)
- **Scaling up**: overcome training data **bottleneck**
  - Use further available English **sense-tagged corpora** (DSO, …)
  - Manual **tagging** (**SenseTagger** application)
  - Exploit word-aligned English-Hungarian **parallel corpora**
  - **Manually** enter disambiguation **rules**

Manual Disambiguation Rules

- Possibility to manually create disambiguation rules for an ambiguous SL item
- Text file format based on WEKA’a arff

```
@item capital-n
@senseid capital_n_to3ke, capital_n_fo3va1ros,
   capital_n_nagybetu3, capital_n_oszlopfo3
@pprior .3 .3 .2 .2
@rules
~ capital_n_to3ke
     go = business
     wo-1 = working
~ capital_n_fo3va1ros
     surf = Capital
     wo+-3 = city
     …
```

Future Work 1.

- Increase disambiguation precision:
  - Closely examine problematic cases (9 items: p. < baseline)
  - Feature engineering:
    - **Optimize** feature subsets for items (Mihalcea, 2003)
    - Feature **weighting**
    - **Filter** feature value-sets (salience)
  - Introduce new contextual features
    - **Syntactic** info (use NP-chunker, shallow parser)
    - **Named Entity** classes (*CITY*, *PERSON*, *COMPANY*, etc.)
    - …
  - Correction of a-priori frequencies (Grozea, 2005)
  - Test other ML learning schemes (SVM, ...)
    - Use larger test set
    - Find optimal features & parameters for algorithm

Future Work 2.

- Scaling up:
  - Use a word-aligned English-Hungarian **parallel corpus** to **automatically** obtain English training instances tagged with Hungarian translations
    - **Hunglish Corpus** (Varga et al `05):
      44m English / 35m Hungarian words
    - Piperidis et al, Specia et al (RANLP-05)
- Verbs, adjectives
  - Verbs: argument structure
- Deal with „subjective factor” of MT end-user
  - Overall precision exceeding baseline not enough: avoid puzzling wrong answers!
  - Estimate disambiguation answer **confidence**;
    if *score* < *threshold*, return **majority sense translation**

Thank you for your attention!