The Web as Collective Mind

Building Large Annotated Data with Web Users' Help

Rada Mihalcea (Univ. of North Texas)

Tim Chklovski (MIT AI lab)

Large Sense-Tagged Corpora Are Needed

- Semantically annotated corpora needed for many tasks
  - Supervised Word Sense Disambiguation
  - Selectional preferences
  - Lexico-semantic relations
  - Topic signatures
  - Subcategorization frames
- Acquisition of linguistic knowledge is one of the main objectives of MEANING
- General “trend”
  - Focus on getting more data
  - As opposed to searching for better learning algorithms

Large Sense-Tagged Corpora Are Needed

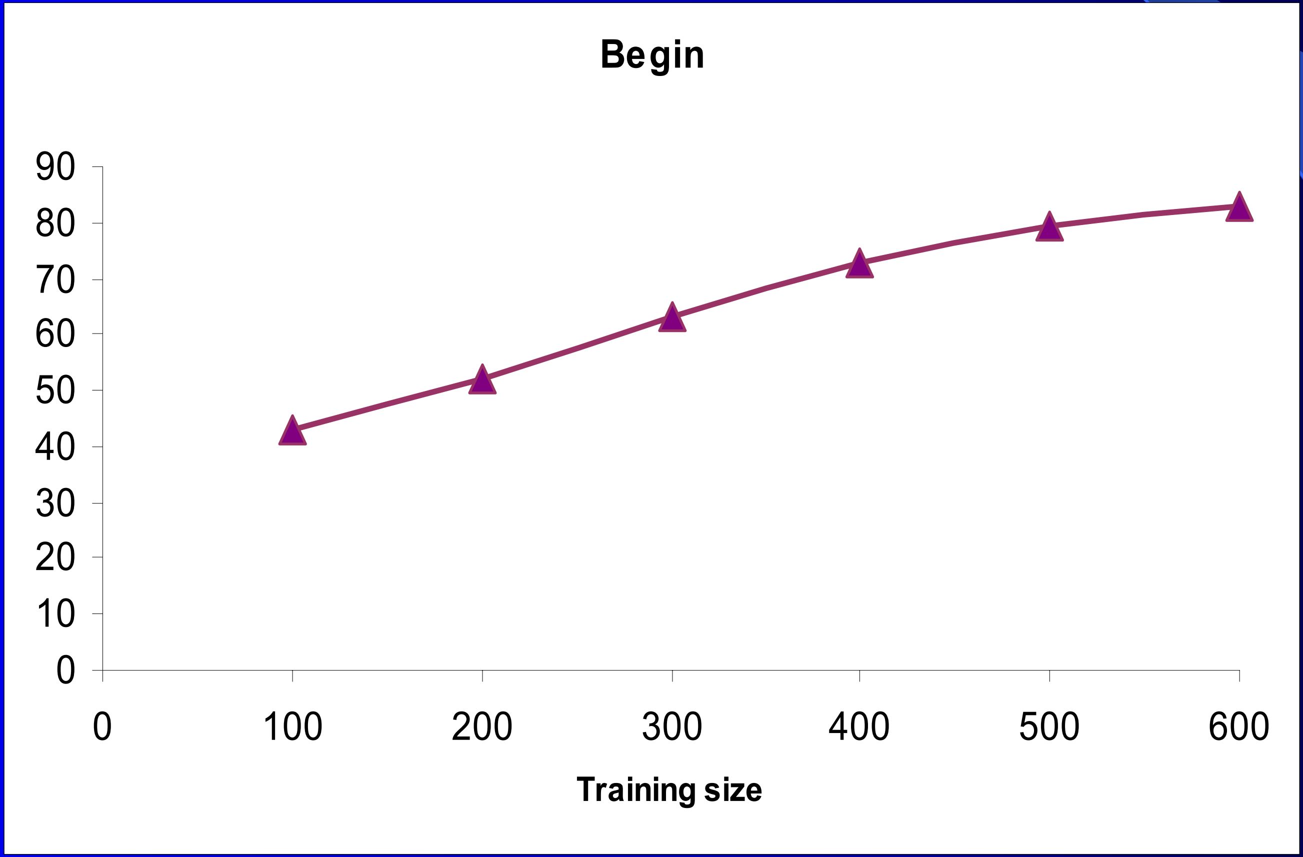
- Large sense-tagged data required for supervised Word Sense Disambiguation
  - Supervised WSD systems have highest performance
  - Mounting evidence that many NLP tasks improve with more data (e.g. Brill, 2001), WSD is no exception
  - Senseval needs training data
    - If we want to see Senseval-5 happening
  - Current method (paid lexicographers) has drawbacks: is expensive and non-trivial to launch and re-launch

How Much Training Corpora ?

begin: a special case in Senseval-2 – data created by mistake!

~700 training examples

~400 test examples



How many ambiguous words?

- English
  - About 20,000 ambiguous words in the common vocabulary (WordNet)
  - About 3,000 high frequency words (H.T. Ng 96)
- Romanian:
  - Some additional 20,000
- Hindi
- French
- ….
- 7,000 different languages!
  - (Scientific American, Aug. 2002)

Size of the problem?

- About 500 examples / ambiguous word
- About 20,000 ambiguous words / language
- About 7,000 languages

dare to do the math…

How much annotated data are available?

- *Line, serve, interest* corpora (2000-4000 instances / word)
- *Senseval-1* and *Senseval-2* data (data for about 100 words, with 75 + 15n examples / word)
- *Semcor* corpus (corpus of 190,000 words, with all words sense-annotated)
- *DSO* corpus (data for about 150 words, with ~500 – 1000 examples / word)

See **senseval.org/data.html** for complete listing

Are we at a dead end?

- Tagging pace with small groups of lexicographers cannot match the data request
- About 16 man-years needed to produce data for about 3,000 English ambiguous words (H.T.Ng)

- **Need to turn towards other, non-traditional approaches for building sense tagged corpora**

Methods for Building Semantically Annotated Corpora

- Automatic acquisition of semantic knowledge from the Web
  - Substitution of words with monosemous equivalents (1999)
  - One of the main lines of experiments in Meaning

Methods for Building Semantically Annotated Corpora

- Bootstrapping
  - Co-training:
    - See over- and under- training issues (Claire Cardie, EMNLP 2001)
  - Iterative assignment of sense labels
    - (Yarowsky 95)
  - Assumes availability of some annotated data to start with

Methods for Building Semantically Annotated Corpora

- Open Mind Word Expert
  - Collect data over the Web
  - Rely on the contribution of thousands of Web users who contribute their knowledge to data annotation
- A different view of the Web

The Web as Collective Mind

Open Mind Word Expert (OMWE)

- Different way to get data: from volunteer contributors on the web
  - Is FREE (assuming bandwidth is free)
  - Part of Open Mind initiative (Stork, 1999)

  - Other Open Mind projects:
    - 1001 Answers
    - CommonSense
    - All available from **http://www.teach-computers.org**

Data / Sense Inventory

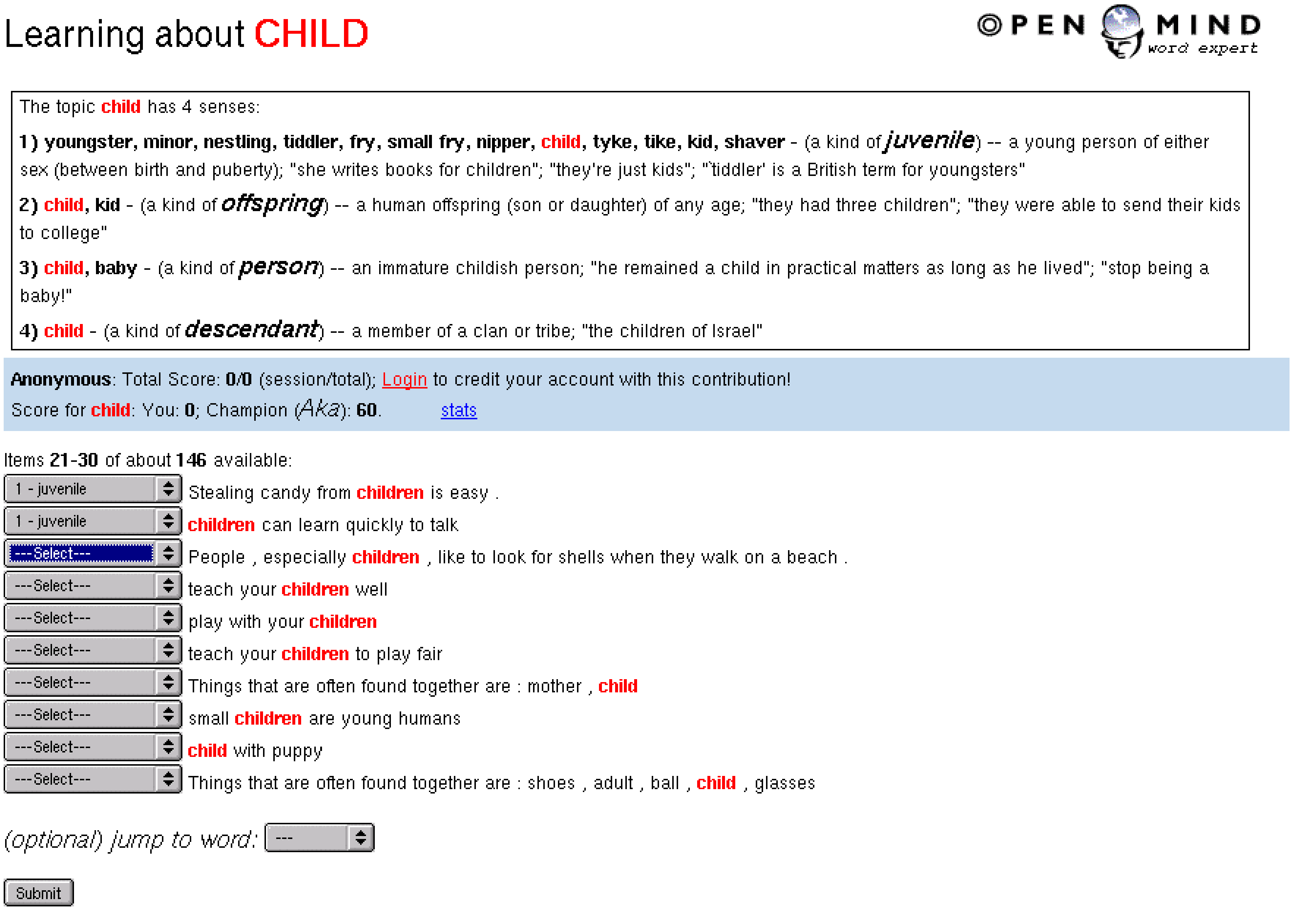
- Uses data from Open Mind Common Sense (Singh, 2002), Penn Treebank, and LA Times (part-of-speech tagged, lemmatized)
- British National Corpus, American National Corpus will be soon added

- WordNet as sense inventory
  - Fine grained
  - Experimenting with clustering based on confusion matrices

Active Learning

- Increased efficiency
- STAFS and COBALT
  - STAFS = semantic tagging using instance based learning with automatic feature selection
  - COBALT = constrained based language tagger
  - STAFS $\cap$ COBALT
    - Agree 54.5% of the times
    - 82.5 / 86.3% precision (fine/coarse senses)

OMWE: http://teach-computers.org



Learning about CHILD

The topic **child** has 4 senses:

1) youngster, minor, nestling, tiddler, fry, small fry, nipper, child, tyke, tike, kid, shaver - (a kind of *juvenile*) -- a young person of either sex (between birth and puberty); "she writes books for children"; "they're just kids"; "`tiddler' is a British term for youngsters"

2) child, kid - (a kind of *offspring*) -- a human offspring (son or daughter) of any age; "they had three children"; "they were able to send their kids to college"

3) child, baby - (a kind of *person*) -- an immature childish person; "he remained a child in practical matters as long as he lived"; "stop being a baby!"

4) child - (a kind of *descendant*) -- a member of a clan or tribe; "the children of Israel"

Anonymous: Total Score: **0/0** (session/total); Login to credit your account with this contribution!

Score for **child**: You: **0**; Champion (*Aka*): **60**. stats

Items **21-30** of about **146** available:

1 - juvenile ⬍ Stealing candy from **children** is easy .

1 - juvenile ⬍ **children** can learn quickly to talk

---Select--- ⬍ People , especially **children** , like to look for shells when they walk on a beach .

---Select--- ⬍ teach your **children** well

---Select--- ⬍ play with your **children**

---Select--- ⬍ teach your **children** to play fair

---Select--- ⬍ Things that are often found together are : mother , **child**

---Select--- ⬍ small **children** are young humans

---Select--- ⬍ **child** with puppy

---Select--- ⬍ Things that are often found together are : shoes , adult , ball , **child** , glasses

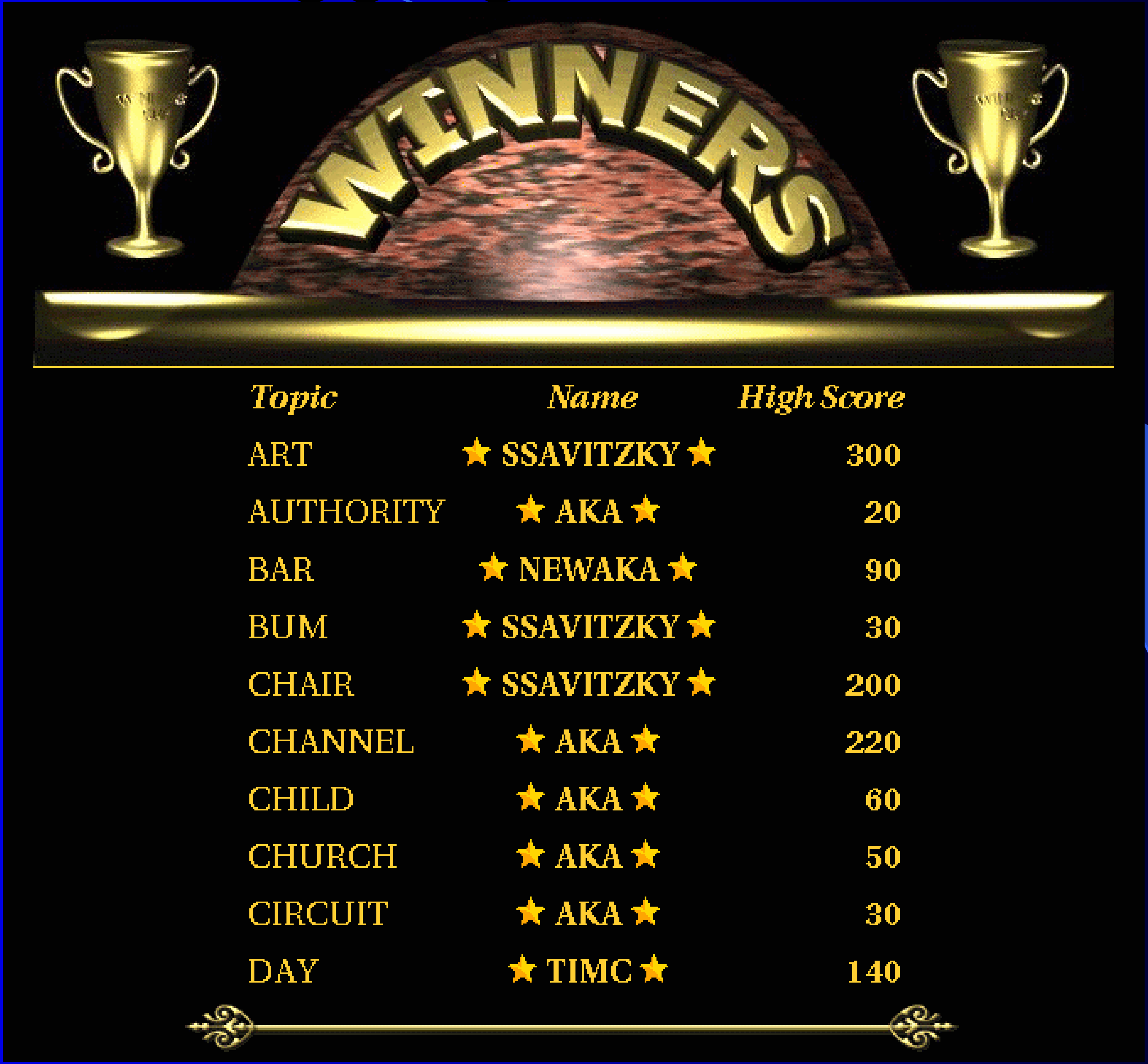
(optional) jump to word: --- ⬍

Submit

Making it Engaging

- Our slogan: “Play a game, make a difference!”
- Can be used as a teaching aid (has special “project” mode):
  - Help introduce students to WSD, lexicography
  - Has been used both at university and high school level
- Features include:
  - Scores, Records, Performance graphs, optional notification when your record has been beaten
  - Prizes
  - Hall of Fame

Tagging for Fame



Topic	Name	High Score
ART	★ SSAVITZKY ★	300
AUTHORITY	★ AKA ★	20
BAR	★ NEWAKA ★	90
BUM	★ SSAVITZKY ★	30
CHAIR	★ SSAVITZKY ★	200
CHANNEL	★ AKA ★	220
CHILD	★ AKA ★	60
CHURCH	★ AKA ★	50
CIRCUIT	★ AKA ★	30
DAY	★ TIMC ★	140

Volume & Quality

- Currently (04/04/2003), about 100,000 tagging acts
- To assure quality, tagging for every item is collected twice, from different users
  - Currently, only perfect agreement cases are admitted into the corpus
  - Preprocessing identifies and tags multi-word expressions (which are the simple cases)
- ITA is comparable with professional tagging:
  - ~67% on first two tags
    - single word tagging collected through OMWE+
    - multi-word tagging automatically performed
  - Kilgarriff reports 66.5% for Senseval-2 nouns on first two tags

INTERESTing Results

- According to Adam Kilgarriff (2000, 2001) replicability is more important than inter-annotator agreement
- A small experiment: re-tag Bruce (1999) “interest” corpus:
  - 2,369 starting examples
  - Eliminate multi-word expressions (about 35% - e.g. “interest rate”) → 1,438 examples
  - 1,066 items with tags that agree → 74% ITA for single words, 83% ITA for entire set
  - 967 items that have a tag identical with Bruce
  - → 90.8% replicability for single words
  - → 94.02% replicability for entire set
  - Kilgarriff (1999) reports 95%

Word Sense Disambiguation using OMWE corpus

- Additional *in-vivo* evaluation of data quality
- Word Sense Disambiguation:
  - STAFS
  - Most frequent sense
  - 10-fold cross validation runs

Word Sense Disambiguation Results

- Intra-corpus experiments: 280 words with data collected through OMWE

Word	Size	MFS	WSD
activity	103	90.00%	90.00%
arm	142	52.50%	80.62%
art	107	30.00%	63.53%
bar	107	61.76%	70.59%
building	114	87.33%	88.67%
cell	126	89.44%	88.33%
chapter	137	68.50%	71.50%
child	105	55.34%	84.67%
circuit	197	31.92%	45.77%
degree	140	71.43%	82.14%
sun	101	63.64%	66.36%
trial	109	87.37%	86.84%

Word Sense Disambiguation Results

Training	Precision		Error rate
examples	baseline	WSD	reduction
any	63.32%	66.23%	9%
> 100	75.88%	80.32%	19%
> 200	63.48%	72.18%	24%
> 300	45.51%	69.15%	43%

The more the better!

- agrees with the conclusions of some of the MEANING experiments
- agrees with previous work (Ng 1997, Brill 2001)

Word Sense Disambiguation Results

- Inter-corpora WSD experiments
- Senseval training data VS. Senseval+OMWE
  - Different sources → different sense distributions

	Senseval		Senseval+OMWE	
art	60.20%	65.30%	61.20%	68.40%
church	62.50%	62.50%	67.20%	67.20%
grip	54.70%	74.50%	62.70%	70.60%
holiday	77.40%	83.90%	77.40%	87.10%
…..				
Average	63.99%	72.27%	64.58%	73.78%

Word Sense Disambiguation Results

- Sense distributions have strong impact on precision
- MEANING experiments
  - 20% difference in precision for data with or without Senseval bias
  - We consider evaluating OMWE data under similar settings (+/- Senseval bias)

Summary of Benefits

- http://teach-computers.org
- A Different View of the Web:

  WWW ≠ large set of pages

  WWW = a way to ask millions of people

  – Particularly suitable for attacking tasks that people find very easy and computers don't
- OMWE approach:
  – Very low cost
  – Large volume (always-on, "active" corpus)
  – Equally High Quality

How OMWE can relate to MEANING efforts?

- Provide starting examples for bootstrapping algorithms
  - Co-training
  - Iterative annotation (Yarowsky 95)
- Provide seeds that can be used in addition to WordNet examples for generation of sense tagged data:
  - Web-based corpus acquisition

A Comparison

	Hand tagging			Open Mind
	with lexicographers	Substitution	Bootstrapping	Word Expert
Automatic	NO	YES	YES-SEMI	NO-SEMI
Human intervention	YES	NO	YES	YES
Expensive?	YES	NO	NO	NO
Time consuming?	YES	NO	SEMI	SEMI
Features: local	YES	NO(?)	YES	YES
Features: global	YES	YES	YES	YES
Uniform coverage?	MAYBE	NO	MAYBE	MAYBE

- Which method to choose?
- The best choice may be a mix!

How MEANING efforts can help our own WSD work?

- Sense tagged data
- Selectional preferences
- Use ExRetrieve to suggest sense labels
  - Speed-up OMWE
  - “clean” ExRetrieve examples
- Cross-validation of (semi)automatic sense labeling experiments

Sneak Preview: OMWE 2.0

- Create data for other languages:
  - Romanian, Hindi, etc.
- Create data for multi-lingual tagging (translations)
  - Multi-lingual tagging
- A slightly improved version of current English OMWE
- Should provide data for three tasks in Senseval-3