CS674 Natural Language Processing

- Last class
  - Word sense disambiguation
    - » Decision lists approach
    - » Weakly supervised
    - » Unsupervised learning
    - » Dictionary-based approaches
- Today
  - Word sense disambiguation
    - » SENSEVAL
  - Noisy channel model
    - » Spelling correction
    - » Pronunciation variation

SENSEVAL-2

- Three tasks
  - Lexical sample
  - All-words
  - Translation
- 12 languages
- Lexicon
  - SENSEVAL-1: from HECTOR corpus
  - SENSEVAL-2: from WordNet 1.7
- 93 systems from 34 teams

Lexical sample task

- Select a sample of words from the lexicon
- Systems must then tag several instances of the sample words in short extracts of text
- SENSEVAL-1: 35 words, 41 tasks
  - 700001 John Dos Passos wrote a poem that talked of \`the <tag>bitter</> beat look, the scorn on the lip."
  - 700002 The beans almost double in size during roasting. Black beans are over roasted and will have a <tag>bitter</> flavour and insufficiently roasted beans are pale and give a colourless, tasteless drink.

Lexical sample task: SENSEVAL-1

Nouns		Verbs		Adjectives		Indeterminates	
-n	N	**-v**	N	**-a**	N	**-p**	N
accident	267	amaze	70	brilliant	229	band	302
behaviour	279	bet	177	deaf	122	bitter	373
bet	274	bother	209	floating	47	hurdle	323
disability	160	bury	201	generous	227	sanction	431
excess	186	calculate	217	giant	97	shake	356
float	75	consume	186	modest	270		
giant	118	derive	216	slight	218		
…	…	…	…	…	…		
TOTAL	2756	TOTAL	2501	TOTAL	1406	TOTAL	1785

All-words task

- Systems must tag almost all of the content words in a sample of running text
  - sense-tag all predicates, nouns that are heads of noun-phrase arguments to those predicates, and adjectives modifying those nouns
  - ~5,000 running words of text
  - ~2,000 sense-tagged words

Translation task

- SENSEVAL-2 task
- Only for Japanese
- word sense is defined according to translation distinction
  - if the head word is translated differently in the given expressional context, then it is treated as constituting a different sense
- word sense disambiguation involves selecting the appropriate English word/phrase/sentence equivalent for a Japanese word

SENSEVAL-2 results

Language	Task	No. of submissions	No. of teams	IAA	Baseline	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40

SENSEVAL plans

- Where next?
  - Supervised ML approaches worked best
    - » Looking at the role of feature selection algorithms
  - Need a well-motivated sense inventory
    - » Inter-annotator agreement went down when moving to WordNet senses
  - Need to tie WSD to real applications
    - » The translation task was a good initial attempt

CS674 Natural Language Processing

- Last class
  - Word sense disambiguation
    - Finish decision lists approach
    - Weakly supervised
    - Unsupervised learning
    - Dictionary-based approaches
- Today
  - Word sense disambiguation
    - SENSEVAL
  - Noisy channel model
    - Spelling correction
    - Pronunciation variation

Correction of spelling errors

- Frequency of spelling errors in human typed text varies from
  - 0.05% of the words in carefully edited newswire, to
  - 38% in difficult applications like telephone directory lookup
- Optical character recognition
  - Higher error rates than human typists
  - Make different kinds of errors, "D"→ "O"; "ri"→"n"
- On-line handwriting recognition

Types of spelling correction

- Non-word error detection
  - Detecting spelling errors that result in non-words
    - *graffe* → *giraffe*
- Isolated-word error correction:
  - Correcting spelling errors that result in non-words
    - Correcting *graffe* to *giraffe*, but looking only at the word in isolation

Kukich, 1992

Types of spelling correction

- Context-dependent error detection and correction
  - Using the context to help detect and correct spelling errors
  - Some of these may accidentally result in an actual word (**real-word errors**)
    - Typographical errors
      - e.g. *there* for *three*
    - Homonym or near-homonym
      - e.g. *dessert* for *desert*, or *piece* for *peace*

Kukich, 1992

Fixing non-word errors

- Detecting non-words
  - Use a dictionary
  - Usually include models of morphology
  - For other types of spelling correction, we'll need a model of spelling variation.

Proposing candidate corrections

- Simplifying assumption: the correct word will differ from the misspelling by a ***single*** insertion, deletion, substitution, or transposition
  - Handles most spelling errors in human typed text
- Generate the candidates by applying any single transformation that results in a word in an on-line dictionary

Candidate corrections for *acress*

		Transformation			
Error	Correction	Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	–	2	deletion
acress	cress	–	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	–	2	5	insertion
acress	acres	–	2	4	insertion

The pronunciation subproblem

[spooky music][music stops]

Head Knight of Ni: Ni!

Knights of Ni: Ni! Ni! Ni! Ni! Ni!

Arthur: Who are you?

Head Knight: We are the Knights Who Say…'Ni'! …

We are the keepers of the sacred words: 'Ni', 'Peng', and 'Neee-wom'!

The pronunciation subproblem

- Given a series of phones, compute the most probable word that generated them.
- Simplifications
  - Given the correct string of phones
    - Speech recognizer relies on probabilistic estimators for each phone, so it's never entirely sure about the identification of any particular phone
  - Given word boundaries
- "I [ni]…"
  - [ni] → *the, neat, need, new, knee, to,* and *you*
  - Based on the (transcribed) Switchboard corpus
- Contextually-induced pronunciation variation

No candidate generation

- Use corpus to expand each pronunciation in advance with all possible variants
- [ni] is stored with the list of words that can generate it

Probabilistic transduction

- surface representation → lexical representation
- sequence of letters in a mis-spelled word → sequence of letters in correctly spelled words
  - *acress* → *actress, cress, acres*
- string of symbols representing the pronunciation of a word in context → string of symbols representing the dictionary pronunciation
  - [er] → *her, were, are, their, your*
  - exacerbated by **pronunciation variation**
    - *the* pronounced as THEE or THUH
    - some aspects of this variation are systematic, like spelling error patterns

Noisy channel model

- Channel introduces noise which makes it hard to recognize the true word.
- **Goal:** build a model of the channel so that we can figure out how it modified the true word…so that we can recover it.

Decoding algorithm

- Special case of **Bayesian inference**
  - Bayesian classification
    - Given observation, determine which of a set of classes it belongs to.
    - Observation
      - string of phones or string of letters
    - Classify into
      - words