"Right from the word go"
identifying MWE for semantic tagging

Paul Rayson

UCREL

Computing Department

Lancaster University

Outline

- Motivation
- Template-based approach
- Statistical approach
- Hybrid methods
- Evaluation
- Conclusion
- Future work

What?

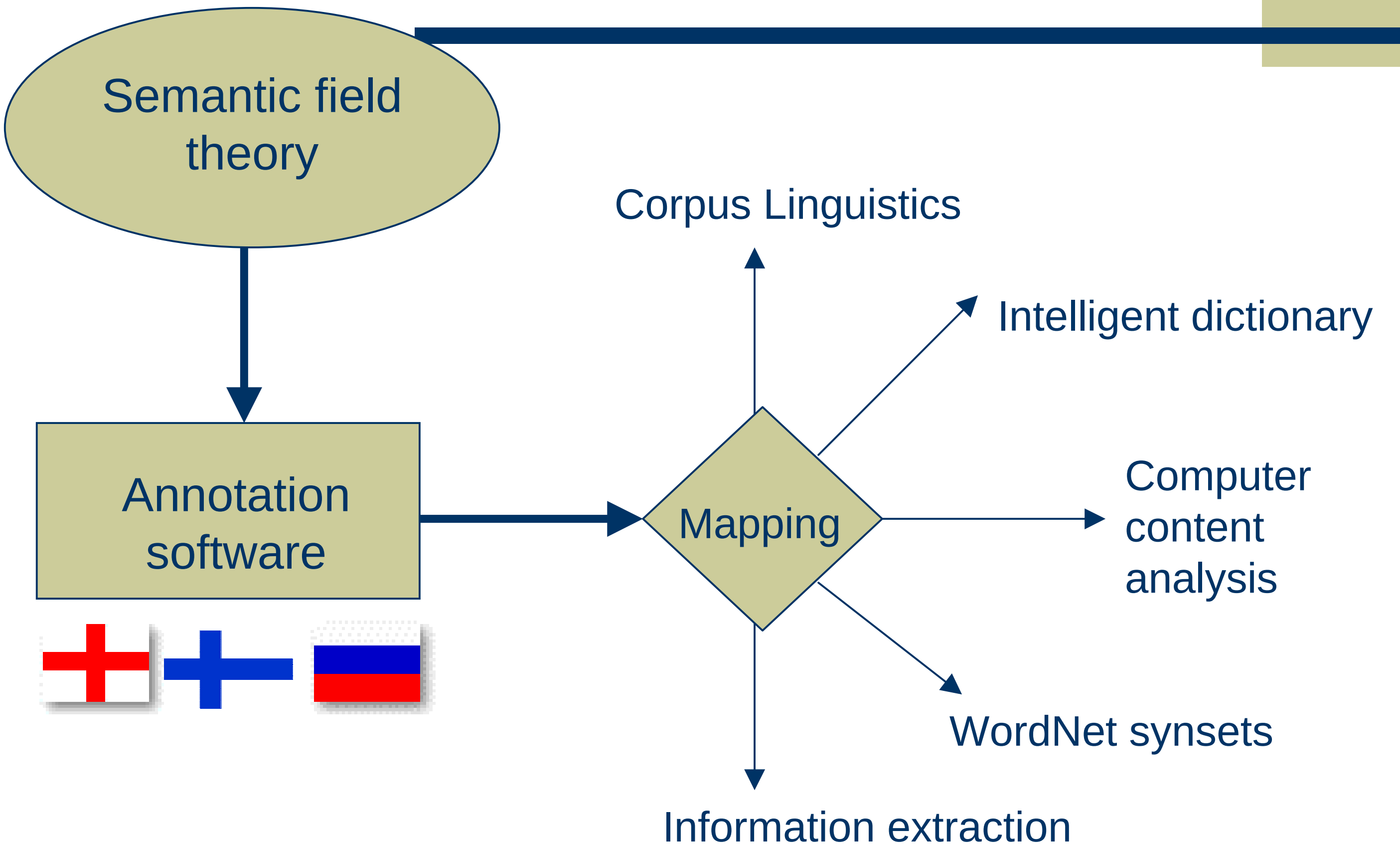
- Lexical bundles
- Colligations
- Collocations
- Prefabricated expressions
- **Multi-word-expressions**
- **Idiomatic expressions**
- **Phrasal verbs**
- **Named entities (people, places, organisations, dates, numbers)**

Why?

- Detecting semantic multi-word-units
- Semantic field annotation
- 16% of words in running text are semantic MWE

Semantic field annotation

Overview


Semantic field theory
Annotation software
Mapping
Corpus Linguistics
Intelligent dictionary
Computer content analysis
WordNet synsets
Information extraction

Application contexts

- Semantic field analysis
- Content analysis
    - Conceptual analysis: USAS, Louw/Nida categories in OpenText.org
    - General category: General Inquirer, Minnesota Contextual Content Analysis
    - Specialised content analysis: RID, Diction
- Market research interview transcript analysis
- Word sense disambiguation: Senseval
- Information extraction / text mining
- Electronic dictionaries

Information extraction

- Requirements reverse engineering to support business process change (Revere)
- Reducing rework through decision management (Tracker)

Links to Lexicography

- The New Intelligent Dictionary (Benedict)
- Providing an interactive user-specified access interface, tailoring the dictionary information supply according to user specifications, incorporating multi-layered entry structure with new information categories and links to corpus data and syntactically- and semantically-based corpus search tools in the dictionary data base.

The task we set ourselves

- Full text tagging, not just selected words
- Tagging the sense in context, not just the word
- Not task specific categories
- Tag set should make sense (psycho)- linguistically
- Flexible category set with hierarchical structure
- Words and multi-word expressions e.g. phrasal verbs (*stubbed out*), noun phrases (*riding boots*), proper names (*United States of America*), true idioms (*living the life of Riley*)

Semantic fields

- AKA conceptual field, a semantic domain, a lexical field, or a lexical domain
- ‘groups together word senses that are related by virtue of their being connected at some level of generality with the same mental concept’
- Not only *synonymy and antonymy* but also *hypernymy and hyponymy*
- E.g. EDUCATION: academic, coaching, coursework, deputy head, exams, PhD, playschool, revision notes, studious, swot, viva

The UCREL Semantic Analysis System

- Hierarchy of 21 major discourse fields expanding into 232 category labels:

Table 1 : The top level of the USAS system

A: General & Abstract Terms	**B:** The Body & the Individual	**C:** Arts & Crafts	**E:** Emotional Actions, States & Processes
F: Food & Farming	**G:** Government & the Public Domain	**H:** Architecture, Building Houses & the Home	**I:** Money & Commerce in Industry
K: Entertainment, Sports & Games	**L:** Life & Living Things	**M:** Movement, Location, Travel & Transport	**N:** Numbers & Measurement
O: Substances, Materials, Objects & Equipment	**P:** Education	**Q:** Linguistic Actions, States & Processes	**S:** Social Actions, States & Processes
T: Time	**W:** The World & Our Environment	**X:** Psychological Actions, States & Processes	**Y:** Science & Technology
Z: Names & Grammatical Words			

Lexical resources

- Lexicon of 51,958 items
  - workshop NN1 I4/H1 P1
- MWE list of 18,808 items
  - travel_NN1 card*_NN* M3/Q1.2
- A small wildcard lexicon
  - *kg NNU N3.5
- A small context rule set of 350 items
  - VB*[Z5] (R*n) (XX) (R*n) V*G*
- Unknown words using WordNet synonym lookup

Main Information and Resources Used

- CLAWS C7 Part-of-speech tagset;
- Single-word lexicon containing POS and possible semantic fields of each word;
- Multiword lexicon and templates containing POS and possible semantic fields of each entry;
- Likelihood ranking of possible semantic fields in the lexicon – mainly subjective process;
- Domain of discourse;
- Contextual information.

MWE Lexicon+Templates of USAS

- It is the main resource for MWE identification
- Sample entries:
  1. table_NN1 tennis_NN1 K5.1
  2. missile_NN1 controller*_NN* G3/S2mf G3
  3. *ing_NN1 machine*_NN* Df/O2
  4. *_* Ocean_N*1 Z2
  5. turn*_* {Np/P*/R*} on_RP A1.1.1 S3.2
  6. smash*_* {Np/P*/R*} to_II {UH/J*} pieces_NN2 A1.1.2

Note: K5.1 – sport; G3 – weapons; S2 – people; df – use the tag of initial word;
O2 – Objects generally; A1.1.1 – general action/making;
A1.1.2 – Damaging & destroying; S3.2 – relationship intimate/sexual;
m – male; f – female; Np – noun phrase.

Five Types of MWE Lexicon Entries

1. Literal MWE list, see sample (1)
2. Allow prefix/suffix changes, see sample (2)
3. Allow words sharing the same prefix/suffix, see sample (3)
4. Allow any preceding/following words, see sample (4)
5. Allowing embedded words, see sample (5) and (6)

Disambiguation of Overlapping MWEs

- Some heuristic rules applied:
  - The longer match is preferred;
  - If the same lengths, the match with fewer embedded words is preferred;
  - More fully-defined match, or the one with fewer wildcards is preferred:
  - Fewer wildcards in the first word of the match;
  - Fewer wildcards in POS tags.

Sample USAS Output

Life_T3/X2.6[i7.2.1 expectancy_T3/X2.6[i7.2.2 was_A3+ poor_I1.1- ,_PUNC the_Z5 average_A6.2+ age_T3 of_Z5 death_L1- was_A3+ 25_T3 **due_A2.2[i8.2.1 to_A2.2[i8.2.2** unhealthy_B2- working_I3.1 conditions_O4.1 and_Z5 Haworth_Z99 's_Z5 diabolical_A5.1-- sanitation_B4 ._PUNC

Note: symbols like [i7.2.1 are MWE tags.

Experiment 1 – USAS for MWE extraction

- Selecting test data;
- Tag the data with USAS and collect MWEs;
- Manually examine the result.

Test Data

- The METER Corpus, built in Sheffield University (Gaizauskas *et al*. 2001), was chosen.
- It is a collection of court reports from PA (British Press Association) and some leading British newspapers.
- The newspaper half of this corpus was drawn as test data.
- Size of the test data: 774 articles containing over 250,000 words.

Why METER Corpus

- It has not been used in USAS training, so good for testing its true capability of MWE extraction.
- A homogeneous corpus with restricted domain, good for extracting domain-specific MWEs.

Issue of Defining MWE

- **A few definitions available, E.g.**
  - **Smadja (1993): recurrent, domain-dependent and cohesive lexical clusters.**
  - **Sag, *et al*. (2001): idiosyncratic interpretations that cross word boundaries.**
  - **Biber *et al*. (2003): lexical bundles that frequently used by many different speakers within a register.**

Which One is Good MWE?

- Experienced disagreements on whether or not a candidate is a good MWE.
- If a candidate can frequently occur in the corpus, it is accepted to be a good MWE.
- Quite a few intuitive/subjective decisions.

Precision of MWE Extraction

Total number of Candidate MWEs extracted = 4,195,

“Good” MWEs found = 3,792,

Precision = 90.39%.

Recall of MWE Extraction

- Estimated based on sample data.
- Randomly selected fifty texts containing 14,711 words.
- Manually checked sample texts to mark-up all good MWEs.

  Results: *Total number of Good MWEs found = 1,511,*

  Good MWEs extracted = 595,

  Recall = 39.38%.

- Given the homogeneous feature of the test corpus, we assume this local recall approximates the global recall of the whole test data.

Precision for Each Semantic Category (1)

Sem field	Total MWEs	Good MWEs	Precision
Z	1,904	1,635	85.87%
T	497	459	92.35%
A	351	328	93.44%
M	254	241	94.88%
N	227	211	92.95%
S	180	177	98.33%
B	131	128	97.71%
G	118	110	93.22%
X	114	104	91.23%
I	74	72	97.30%
Q	67	63	94.03%
E	58	53	91.38%
H	53	52	98.11%
K	48	45	93.75%
P	39	37	94.87%
O	32	29	90.63%
F	24	24	100.00%
L	11	11	100.00%
Y	6	6	100.00%
C	5	5	100.00%
W	2	2	100.00%
Total	4,195	3,792	90.39%

Precision for Each Semantic Category (2)

- **Precisions for individual categories range between 91.23% to 100%.**
- **Categories F (*food & farming*), L (*life & living things*), Y (*science & technology*), C (*arts & crafts*), W (*the world & environment*) obtain 100%, but fewer MWEs as well.**
- **Category Z (*names & grammatical words*), containing 45.39% of the MWEs extracted, obtains the lowest precision (85.87%).**
- **Many word pairs are tagged as names by mistake.**

Precisions for MWEs of Different Lengths

MWE length	Total MWEs	Good MWEs	Precision
2	3,378	3,105	91.92%
3	700	575	82.14%
4	95	91	95.44%
5	18	17	94.44%
6	4	4	100.00%
Total	4,195	3,792	90.39%

- **More short MWEs than longer ones.**
- **Generally better precision for longer MWE.**
- **Typical tri-gram errors: many *CIW*+prep.+*CIW* structures are tagged as geographical names by mistake,**

  e.g. *Sunday_on_United*, *Tanzania_on_August*, etc.

Note: CIW – capital initial word

Precisions for MWEs of Different Frequencies

Frequency	Total MWEs	Good MWEs	Precision
1	2,164	1,892	87.43%
2	750	695	92.67%
3 - 4	616	570	92.53%
5 - 7	357	345	96.64%
8 - 20	253	238	94.07%
21 - 117	55	52	94.55%
Total	4,195	3,792	90.39%

- **Generally, slightly better precisions for more frequent MWEs.**
- **Successfully extracted MWEs of low frequencies – 69.46% and 68.22% of the extracted MWEs and accepted MWEs occur only once or twice.**

Experiment 2: A Collocation-based Statistical Algorithm for MWE Extraction

- **Algorithm:**
  - Pos-tag the text using CLAWS POS tagger;
  - Collect collocates using the co-occurrence association score;
  - Using the collection of collocates as a statistical dictionary, check the affinity between closely adjacent words to create affinity distribution map;
  - Based on the affinity distribution, collect the word clusters (not just word pairs) that are subject to relatively stronger affinity.
  - Optionally, apply simple linguistic filters to remove frequent errors (not used in this experiment).

Log-likelihood Score

Contingency Table:

Suppose X and Y are a pair of words,

- ***a – number of windows in which X and Y co-occur,***
- ***b – number of windows in which only X occurs,***
- ***c – number of windows in which only Y occurs,***
- ***d – number of windows in which none of them occurs,***

then

$$G2 = 2\,(a\ln a + b\ln b + c\ln c + d\ln d - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d)) + (a+b+c+d)\ln(a+b+c+d))$$

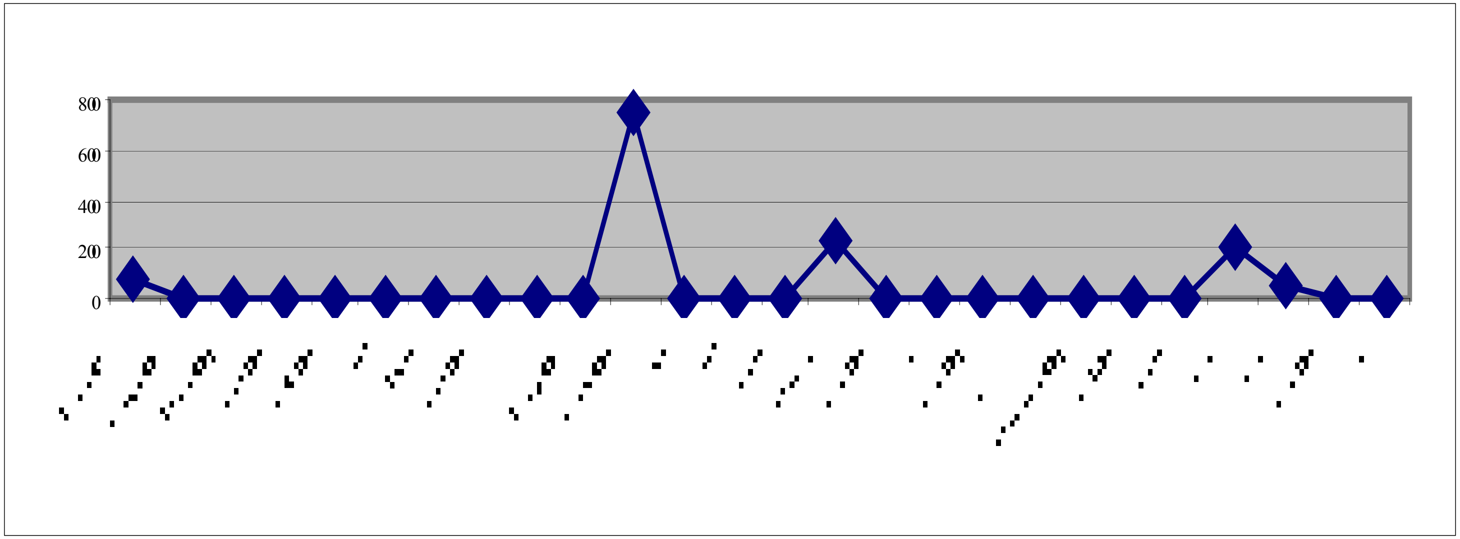
Filter *of t*-score

t-score is used for filtering out some insignificant word collocations:

$$t = \frac{prob(W_a, W_b) - prob(W_a)\,prob(W_b)}{\sqrt{\frac{1}{M}\,prob(W_a, W_b)}}$$

Affinity Distribution of A Sample Sentence

Deputy_NN1 principal_NN1 Alden_NN1 was_VBDZ jailed_VVN for_IF 15_MC years_NNT2 after_II being_VBG found_VVN guilty_JJ of_IO five_MC indecent_JJ assaults_NN2 ,_, one_MC1 gross_NNO indecency_NN1 and_CC four_MC serious_JJ sexual_JJ assaults_NN2 ._.



MWE Marked Output

- <s><mwe> ***Deputy_NN1 principal_NN1*** **</mwe> Alden_NN1 was_VBDZ jailed_VVN for_IF 15_MC years_NNT2 after_II being_VBG <mwe>** ***found_VVN guilty_JJ*** **</mwe> of_IO five_MC <mwe>** ***indecent_JJ assaults_NN2*** **</mwe> ,_, one_MC1 gross_NNO indecency_NN1 and_CC four_MC <mwe>** ***serious_JJ sexual_JJ assaults_NN2*** **</mwe> ._.</s>**

Overall Evaluation in Comparison to USAS

Statistical Tool: ***Number of Candidates = 3,306***
Accepted MWEs = 2,705
Precision = 81.85%

Tools	MWEs	Precision	Recall
Semantic tagger	3,792	90.39%	39.38%
Statistical tool	2,705	81.85%	22.70%

Comparative MWE Frequency Distributions

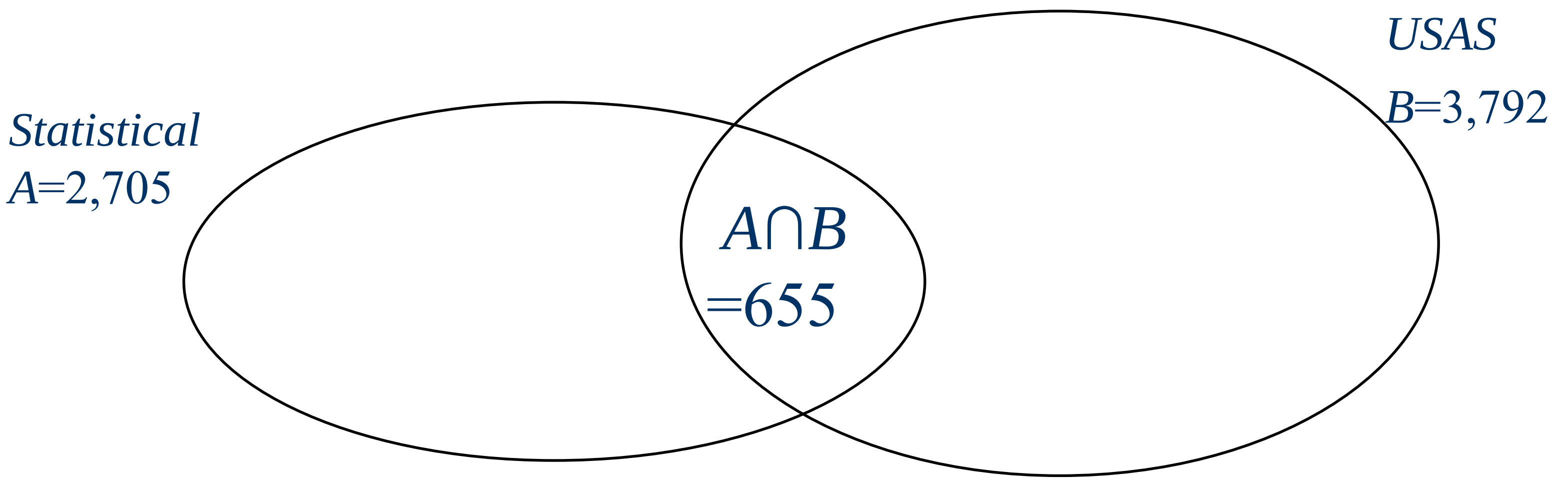
MWE freq	Semantic tagger	Percen-tage	Statistical tool	Percen-tage
1	1,892	49.89%	402	14.86%
2	695	18.33%	274	10.13%
3 - 4	570	15.03%	1,216	44.95%
5 - 7	345	9.10%	504	18.63%
8 – 20	238	6.28%	261	9.65%
>= 21	52	1.37%	48	1.77%
Total	3,792	100.00%	2,705	100.00%

Comparative MWE Length Distributions

MWE length	Semantic tagger	Percen-tage	Statistical tool	Percen-tage
2	3,105	81.88%	2,046	75.64%
3	575	15.16%	494	18.26%
4	91	2.40%	121	4.47%
5	17	0.45%	39	1.44%
>= 6	4	0.11%	5	0.18%
Total	3,792	100.00%	2,705	100.00%

Overlap of MWEs Extracted by Two Approaches

Observation: 75.79% and 82.73% of the MWEs extracted by USAS and statistical tool are complementary results.



Combine Two Approaches Together

Number of MWEs Extracted = 5,842

Precision = 88.14%

Recall = 50.5%

Conclusion

- Implications:
  - USAS provides a practical tool for MWE extraction - not only extract MWEs, but also their semantic field information.
  - As a symbolic tool, it doesn’t know guessing ---

    I only know what I am told.
  - A statistical tool can efficiently extract frequent domain-specific MWEs, but less efficient in identifying low-frequency MWEs
  - We observed that semantic tagger and the statistical tool are complementary for NEW extraction.
  - We suggest that MWE extraction can be significantly improved by combining symbolic tools and statistical tools.

Ongoing work

- Extraction of MWU from EFL corpora
- Semantic field taggers for Finnish and Russian

Future work

- Classification task
- Lemma templates
- Identification of figurative expressions

Questions?

- Further information at http://www.comp.lancs.ac.uk/ucrel/usas/
- Scott Songlin Piao, Paul Rayson, Dawn Archer and Tony McEnery (2005). Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE Extraction.

  Computer Speech and Language.

Appendix

Disambiguation methods (1)

- 1. POS tag

spring	temporal noun	[season sense]
spring	common noun	[coil sense] [water source sense]
spring	verb	[jump sense]

- 2. General likelihood ranking for single-word and MWE tags
  - *green* referring to [colour] is generally more frequent than *green* meaning [inexperienced]
- 3. Overlapping MWE resolution
  - Heuristics applied: semantic MWEs override single word tagging, length and span of MWE also significant

Disambiguation methods (2)

- 4. Domain of discourse
  - adjective *battered*
    - [Violence] (e.g. battered wife)
    - [Judgement of Appearance] (e.g. battered car)
    - [Food] (e.g. battered cod)
- 5. Text-based disambiguation
  - one sense per text
- 6. Context rules
  - *Auxiliary verbs (be/do/have)*
  - *account* of NP [narrative]
  - balance of xxx *account* [financial]

Disambiguation methods (3)

- 7. Local probabilistic
  - *account* occurring in the company of *financial*, *bank*, *overdrawn*, *money*
  - surrounding words, POS tags or semantic fields
  - span of words
  - co-occurrence measures rather than HMM