# Lexical Cognitive Networks with Application to Semantic Similarity Computation and Affective Text Analysis

Alexandros Potamianos

Dept. of ECE, Technical Univ. of Crete, Chania, Greece

# Acknowledgements


- Elias Iosif: Semantic similarity computation, semantic networks
- Nikos Malandrakis: Affective models for text and multimedia
- Shri Narayanan (USC): Affective modeling of dialogue interaction


## References


[1] E. Iosif and A. Potamianos. 2010. “Unsupervised semantic similarity computation between terms using web documents”. IEEE Transactions on Knowledge and Data Engineering.
[2] N. Malandrakis, A. Potamianos, E. Iosif, S. Narayanan. 2011. “Kernel methods for affective lexicon creation”. Proc. Interspeech.
[3] — . 2011. “EmotiWord: Affective Lexicon Creation with Application to Interaction and Multimedia Data”. Proc. of MUSCLE workshop.
[4] E. Iosif and A. Potamianos. 2012. “Semsim: Resources for normalized semantic similarity computation using lexical networks”. In Proc. LREC.
[5] N. Malandrakis, E. Iosif, A. Potamianos. 2012. “DeepPurple: Estimating Sentence Semantic Similarity using N-gram Regression Models and Web Snippets”. In Proc SemEval (collocated with NAACL-HLT).
[6] E. Iosif and A. Potamianos. 2012. “Similarity computation using semantic networks created from web-harvested data”. Natural Language Engineering (submitted to).

# Problem Definition

- Semantic Similarity Computation
  - Given a pair of words or terms $(w_i, w_j)$
  - Compute semantic similarity between them $S(i, j)$
- Related tasks
  - Phrase or sentence level semantic similarity
  - Strength of associative relation between words
  - Affective score (valence) of words and sentences
- Motivation
  - Organizing principle of human cognition
  - Building block of machine learning in NLP/semantic web
  - Entry point for the semantics of language

# Problem Definition

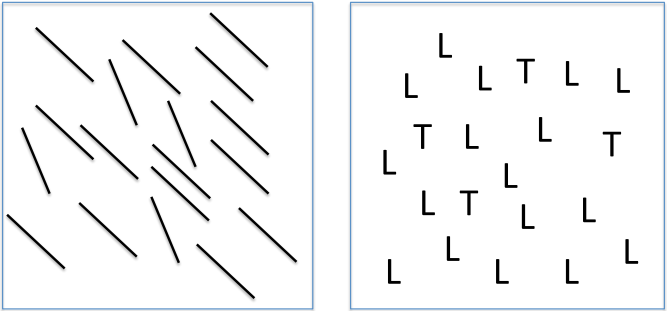
- How to define semantic relations between words?
  - *Linguists*: well-defined in terms of a handful of relations
  - *Cognitive scientists*: no idea!
- How to define associative relations between words?
  - *Linguists*: long-tail of relations, why is this thing useful anyhow?
  - *Cognitive scientists*: well-defined experimentally via priming
- What is lexical priming?
  - Activation of associated words in cognitive lexical net
  - Cache for quick access of most probable candidates
  - Look-ahead useful for pruning improbable hypotheses

# System 1 vs System 2

- Using Kahneman's (and others) formalism:
  - System 1 (intuition): generates
    - impressions, feelings, and inclinations
  - System 2 (reason): turns System 1 input into
    - beliefs, attitudes, and intentions
- Associative relations reside in System 1
- But where do semantic relations reside?

# Example

- Example from vision: system 1 vs system 2



# Semantic Relationship Scoring

- Experiment
  - Please quickly rate the following word pair in terms of semantic similarity using a score between 0 (totally dissimilar) and 4 (semantically equivalent). Record this score.
  - Then take your time adjusting this score to the most appropriate semantic similarity value between 0 and 4.

# Semantic Relationship Scoring

- Experiment
  - Please quickly rate the following word pair in terms of semantic similarity using a score between 0 (totally dissimilar) and 4 (semantically equivalent). Record this score.
  - Then take your time adjusting this score to the most appropriate semantic similarity value between 0 and 4.
- Word pair: (hand, glove)

# Semantic Relationship Scoring

- Experiment
  - Please quickly rate the following word pair in terms of semantic similarity using a score between 0 (totally dissimilar) and 4 (semantically equivalent). Record this score.
  - Then take your time adjusting this score to the most appropriate semantic similarity value between 0 and 4.
- Word pair: (hand, glove)
- Results:
  - System 1: strong association fast score 3 or 4
  - System 2: weak (or non-existent) semantic relationship slow score 2 or 3

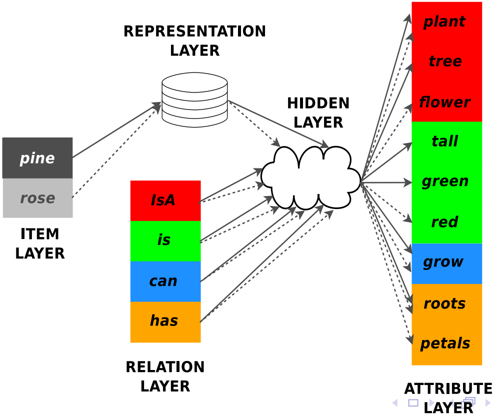
# Associative Anchoring

- What you experienced: associative anchoring
- Anchoring is a cognitive deficiency due to system 1 vs 2 cognitive organization, e.g.,
  - $x = 1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7$ vs
  - $y = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$
  - Fast estimate of $y$ greater than that of $x$
  - Slow estimate of $y$ greater than that of $x$
- Semantic score of (hand, glove) should be 0 or 1
  - instead due to associative anchoring 2 or 3

Semantic score a (system 2) post-correction of association (system 1) score

# A cognitive model of lexical semantics

- Distributed representation *[Rogers & McClelland, '04]*
- Similarity: common vs. distinctive attributes *[Tversky, '77]*

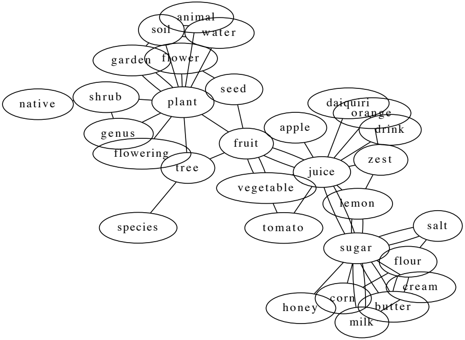


# Main approaches of lexical semantics

- Word are associated with feature vectors
  - crisp, parsimonious representation of semantics
- Distributional semantic models (DSMs)
  - Semantic information extracted from word frequencies
  - Estimate co-occurence counts of word pairs or triplets
  - Estimate statistics of word context vectors
- Semantic networks
  - discovery of new relations via systematic co-variation
  - robust estimates – smoothing corpus statistics over network
  - rapid language acquisition

# Example of Semantic Network

- Linked nodes: lexicalized senses and attributes
  - Informative for semantic similarity computation
- Computation of structural properties, e.g., cliques



# Proposed semantic similarity two-tier system

- Unifies the three approaches
- Fuzzy vs explicit semantic relations
- Word senses vs words vs concepts
- A two tier system
  - An associative network backbone
  - Semantic relations defined as operations on network neighborhoods (cliques)
- Consistent with system 1 vs system 2 view
- Furthermore we believe that the
  - underlying network consists of word senses, and
  - is a low dimensional semi-metric space

# Semantic Similarity Estimation by Machines

- Resource-based, e.g., WordNet
  - Require expert knowledge
  - Not available for all languages
- Corpus-based
  - Distributional semantic models (DSMs)
  - Unstructured (unsupervised): no use of linguistic structure
  - Structured: use of linguistic structure
  - Pattern-based, e.g., Hearst patterns
- Mixed

# Semantic Sim. Computation: Sense Similarity

- Maximum sense similarity assumption *[Resnik, '95]*:
  - Similarity of words equal to similarity of their closest senses
  - If words are considered as sets of word senses, this is the "common sense" set distance
- Given words $w_1$, $w_2$ with senses $s_{1i}$, $s_{2j}$

$$S(w_1, w_2) = \max_{ij} S(s_{1i}, s_{2j})$$

# Semantic Sim. Computation: Attributional Similarity

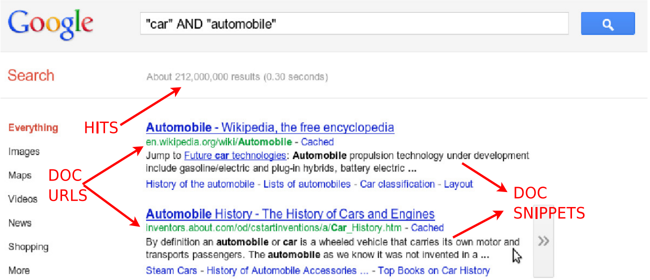
Attributional similarity assumption

- Attributes (features) reflect semantics
  - Item-Relation-Attribute, e.g., canary-color-yellow
- Main representation schemes
  - Hierarchical/Categorical
    - Mainly taxonomic relations, e.g., IsA, PartOf
  - Distributed (networks)
    - Open set of relations, e.g., Cause-Effect, etc
- Similarity between words
  - Function of attribute similarity
  - Defined wrt representation

# Types of Similarity Metrics

- Co-occurrence-based
  - Assumption: co-occurrence implies relatedness
  - Co-occurrence counts: web hits, corpus-based
  - Examples: Dice coef., point-wise mutual information, ...
- Context-based
  - Assumption: context similarity implies relatedness (distributional hypothesis of meaning)
  - Contextual features extracted from corpus
  - Examples: Kullback-Leibler divergence, cosine similarity, ...
- Network-based (proposed)
  - Build lexical net using co-occurrence and/or context sim.
  - Notion of semantic neighborhoods
  - Assumptions: neighborhoods capture word semantics

# Queries to Web Search Engines




- Number of hits
- Document URLs (download)
- Document snippets

# Corpus Creation using Web Queries

- Two types of web queries
  - AND, e.g., "money + bank"
    "... leading **bank** in India offering online **money** transfer ..."
  - IND, e.g., "bank"
    "... downstream parallel to the **banks** of the river ..."
- AND queries
  - Pros: Similarity computation highly correlated (0.88) with human ratings *[Iosif & Potamianos, '10]*
  - Cons: Quadratic query complexity wrt lexicon $L$
- IND queries
  - Pros: Linear query complexity wrt lexicon $L$
  - Cons: Sense ambiguity: moderate correlation (0.55)

# Semantic Similarity Estimation

- Co-occurence based metrics
  - From web: hits of IND, AND queries
  - From (web) corpus: co-occurence counts at the snippet or sentence level
  - Metrics: Dice, Jacard, Mutual Information, Google
- Context-based metrics
  - Download a corpus of documents of snippets using IND queries
  - Construct lexical context vector for each word (window $\pm 1$)
  - Cosine similarity using binary or log-weighted counts

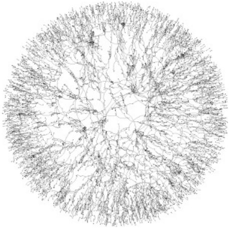
# Enter semantic networks

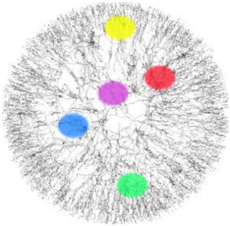
- Why do IND queries fail to achieve good performance?
  1. Word senses are often semantically diverse
     - co-occurrence acts as a semantic filter
  2. Word senses have poor coverage in IND queries
     - rare word senses of words not well-represented
- Solution: use semantic networks
  1. Create a corpus for all words in lexicon (not just semantic similarity pair)
  2. Use semantic neighborhoods for semantic cohesion
     - improved robustness
  3. Inverse frequency word-sense discovery
     - discover rare senses via co-occurrence with infrequent words

# Corpus and Network Creation

- Goals
  - Linear web query complexity for corpus creation
  - New similarity metrics with high performance
- Proposed method
  - IND queries to aggregate data for large $L$ ( $\approx 9K$ )
  - Create network and semantic neighborhoods
  - Neighborhood-based similarity metrics
- Advantages
  - Network: parsimonious representation of corpus statistics
  - Smooth distributions
  - Rare words: well-represented
  - Enable discovery of less frequent senses

# Lexical Network - Semantic Neighborhoods





## Lexical Network

- Undirected graph $G = (N, E)$
  - Vertices $N$: words in lexicon $L$
  - Edges $E$: word similarities

## Semantic Neighborhoods

- For word $i$ create subgraph $G_i$
- Select neighbors of $i$
  - Compute $S(i, j), \forall j \in L, i \neq j$
  - Sort $j$ according to $S(i, j)$
  - Select $|N_i|$ top-ranked $j$

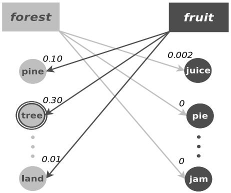
# Semantic Neighborhoods: Examples

| Word | Neighbors |
|---|---|
| automobile | **auto**, truck, **vehicle**, **car**, **engine**, bus, . . . |
| car | truck, **vehicle**, travel, service, **price**, **industry**, . . . |
| slave | slavery, beggar, **nationalism**, society, **democracy**, **aristocracy**, . . . |
| journey | **trip**, holiday, **culture**, **travel**, **discovery**, **quest**, . . . |

- Synonymy
- Taxonomic: IsA, Meronymy
- Associative
- Broader semantics/pragmatics
- . . .

# Neighborhood-based Similarity Metrics: $M_n$

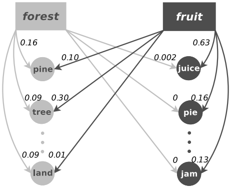
$M_n$ metric: maximum similarity of neighborhoods




- Motivated by maximum sense similarity assumption
  - Neighbors are semantic features denoting senses
  - Similarity of two closest senses
- Select max. similarity: $M_n(\text{"forest"}, \text{"fruit"}) = 0.30$

# Neighborhood-based Similarity Metrics: $R_n$

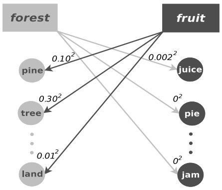
$R_n$ metric: correlation of neighborhood similarities




- Motivated by attributional similarity assumption
  - Neighborhoods encode word attributes (or features)
  - Similar words have co-varying sim. wrt their neighbors
- Compute correlation $r$ of neighborhood similarities
  - $r_1((0.16 \ldots 0.09), (0.10 \ldots 0.01)), r_2((0.002 \ldots 0), (0.63 \ldots 0.13))$
- Select max. correlation: $R_n(\text{“forest”}, \text{“fruit”}) = -0.04$

# Neighborhood-based Similarity Metrics: metric $E_n^{\theta=2}$

$E_n^{\theta=2}$ metric : sum of squared neighborhood similarities




- Motivation: middle road between $M_n$ and $R_n$
  - Accumulation of word–to–neighbor similarities
  - Non-linear weighting of similarities via $\theta = 2$
- $E_n^{\theta=2}$(“forest”, “fruit”)= $\sqrt{(0.10^2 + \cdots + 0.01^2) + (0.002^2 + \cdots + 0^2)} = 0.22$

# Performance of net-based similarity metrics

- Task: similarity judgment on noun pairs
- Dataset: MC *[Miller and Charles, 1998]*
- Evaluation metric: Pearson's correlation wrt to human ratings

| Dataset | Neighbor selection | Similarity computation | Metrics | | |
|---|---|---|---|---|---|
| | | | $M_{n=100}$ | $R_{n=100}$ | $E^{\theta=2}_{n=100}$ |
| MC | co-occur. | co-occur. | 0.90 | 0.72 | **0.90** |
| MC | co-occur. | context | **0.91** | 0.28 | 0.46 |
| MC | context | co-occur. | 0.52 | **0.78** | 0.56 |
| MC | context | context | 0.51 | 0.77 | 0.29 |

# Main findings

- Network construction
  - Co-occurence metrics achieve high-recall for word senses
  - Context-based metrics achieve high-recall for attributes
- Semantic similarity performance
  - Co-occurence a more robust feature that context
  - Max sense similarity assumption is valid and gives best performance
  - Attributional similarity assumption valid for certain cases/languages

# Performance of web-based similarity metrics

- For MC dataset

| Feature | Description | Correlation |
|---|---|---|
| context | AND queries | 0.88 |
| context | IND queries | 0.55 |
| context | IND queries: network | 0.90 |

- Comparable to structured DSMs, WordNet-based approaches

# Minimum Error Semantic Similarity

- Assumption 1

  Senses lexicalized as single words

- Assumption 2

  Sim. of $w_i$, $w_j$: pairwise max. sim. between their senses

- Assumption 3

  3a. $w_i$, $w_j$ always co-occur with their two closest senses

  3b. ...

- Assumption 4

  4a. Uniform distribution of senses

  4b. ...

# Performance of min error semantic similarity

- Modify pointwise mutual info. $I(w_i, w_j) = \log \frac{\hat{p}(w_i,w_j)}{\hat{p}(w_i)\hat{p}(w_j)}$ as

$$I_\alpha(w_i, w_j) = \frac{1}{2}\left[\log \frac{\hat{p}(w_i, w_j)}{\hat{p}^\alpha(w_i)\hat{p}(w_j)} + \log \frac{\hat{p}(w_i, w_j)}{\hat{p}(w_i)\hat{p}^\alpha(w_j)}\right]$$

- Assumptions: 1, 2, 3a, and 4b
- Co-occurrence considered at sentence-level
- $\alpha$ estimated to max. sense coverage of sem. neigh.
- Task: similarity judgment, correlation wrt to human ratings

| Dataset | $I$ | $I_\alpha$ |
|---|---|---|
| MC | 0.78 | **0.89** |
| RG | 0.77 | **0.84** |
| WS353 | 0.60 | **0.68** |

# SemEval 2012: Sentence Level Semantic Similarity

- BLEU-based semantic similarity metric:
  - Baseline BLEU: using single BLEU hit rate as rating
  - Semantic Similarity (SS) BLEU: modified unigram BLEU that includes semantic similarity of non-matched words

Correlation performance of 1-gram BLEU scores with semantic similarity metrics (nouns-only)

| | par | vid | euro | Mean | Ovrl |
|---|---|---|---|---|---|
| BLEU | 0.54 | 0.60 | 0.39 | 0.51 | 0.58 |
| SS-BLEU WordNet | 0.56 | **0.64** | **0.41** | **0.54** | 0.58 |
| SS-BLEU $I(i,j)$ | 0.56 | 0.63 | 0.39 | 0.53 | **0.59** |
| SS-BLEU $I_a(i,j)$ | **0.57** | **0.64** | 0.40 | **0.54** | 0.58 |

# Contributions

Proposed a language agnostic, unsupervised and scalable algorithm for semantic similarity computation

- No linguistic knowledge required, works from text corpus or using a web query engine
- Shown to perform at least as well as resource-based semantic similarity computation algorithms, e.g., WordNet-based methods

# EmotiWord: Affective Lexicon Creation with Application to Interaction and Multimedia Data

# Motivation

- Affective text labeling at the core of many multimedia applications, e.g.,
  - Sentiment analysis
  - Spoken dialogue systems
  - Emotion tracking of multimedia content
- Affective lexicon is the main resource used to bootstrap affective text labeling
  - Lexica are currently of limited scope and quality

# Goals and Contributions

Our goal: assigning continuous high-quality polarity ratings to any lexical unit

- We present a method of expanding an affective lexicon, using web-based semantic similarity
- Assumption: semantic similarity implies affective similarity.
- The expanded lexica are accurate and broad in scope, e.g., they can contain proper nouns, multi-word terms

# Our lexicon expansion method

Expansion of [Turney and Littman, '02].
Assumption: the valence of a word can be expressed as a linear combination of its semantic similarities to a set of seed words and their valence ratings:

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i \, v(w_i) \, d(w_i, w_j), \tag{1}$$

- $w_j$ : the wanted word
- $w_1 \ldots w_N$ : seed words
- $v(w_i)$ : valence rating of word $w_i$
- $a_i$ : weight assigned to seed $w_i$
- $d(w_i, w_j)$ : measure of semantic similarity between words $w_i$ and $w_j$

Given

- an initial lexicon of $K$ words
- a set of $N < K$ seed words

we can use (1) to create a system of $K$ linear equations with $N + 1$ unknown variables:

$$\begin{bmatrix} 1 & d(w_1, w_1)v(w_1) & \cdots & d(w_1, w_N)v(w_N) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & d(w_K, w_1)v(w_1) & \cdots & d(w_K, w_N)v(w_N) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} 1 \\ v(w_1) \\ \vdots \\ v(w_K) \end{bmatrix} \quad (2)$$

Solving with Least Mean Squares estimation provides the weights $a_i$.

# Example, $N = 10$ seeds

| Order | $w_i$ | $v(w_i)$ | $a_i$ | $v(w_i) \times a_i$ |
|---|---|---|---|---|
| 1 | mutilate | -0.8 | 0.75 | -0.60 |
| 2 | intimate | 0.65 | 3.74 | 2.43 |
| 3 | poison | -0.76 | 5.15 | -3.91 |
| 4 | bankrupt | -0.75 | 5.94 | -4.46 |
| 5 | passion | 0.76 | 4.77 | 3.63 |
| 6 | misery | -0.77 | 8.05 | -6.20 |
| 7 | joyful | 0.81 | 6.4 | 5.18 |
| 8 | optimism | 0.49 | 7.14 | 3.50 |
| 9 | loneliness | -0.85 | 3.08 | -2.62 |
| 10 | orgasm | 0.83 | 2.16 | 1.79 |
| - | $w_0$ *(offset)* | 1 | 0.28 | 0.28 |

# Sentence Tagging

Simple combinations of word ratings:

- linear (average)

$$v_1(s) = \frac{1}{N}\sum_{i=1}^{N} v(w_i)$$

- weighted average

$$v_2(s) = \frac{1}{\sum_{i=1}^{N} |v(w_i)|}\sum_{i=1}^{N} v(w_i)^2 \cdot \text{sign}(v(w_i))$$

- max

$$v_3(s) = \max_i(|v(w_i)|) \cdot \text{sign}(v(w_z)), \quad z = \arg\max_i(|v(w_i)|)$$

# N-gram Affective Models

- Generalize method to n-grams

$$v_i(s) = a_0 + a_1 v_i(unigram) + a_2 v_i(bigram)$$

- Starting from all 1-grams and 2-grams, select terms:
  1. Backoff: use overlapping bigrams as default, revert to unigrams based on mutual information-based criterion
  2. Weighted interpolation: use all unigrams and bigrams as default, reject bigrams based on criterion
- In both cases unigrams and bigrams are given linear weights, trained using LMS

# Evaluation

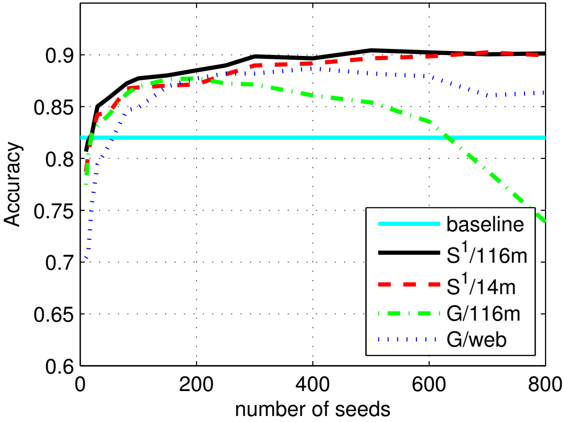
- ANEW Word Polarity Detection Task
  - Affective norms for English words (ANEW) corpus
  - 1.034 English words, continuous valence ratings
- General Inquirer Word Polarity Detection
  - General Inquirer words corpus
  - 3.607 English words, binary valence ratings
- BAWLR Word Polarity Detection Task
  - Berlin affective word list reloaded (BAWLR) corpus
  - 2.902 German words, continuous valence ratings
- SemEval 2007 Sentence Polarity Detection
  - SemEval 2007 News Headlines corpus
  - 1.000 English sentences, continuous valence ratings
  - ANEW used for lexicon training
  - 250 sentence development set used for word fusion training

# Experimental Procedure

- Corpus selection
  - Web corpus (web)
  - Lexically balanced web corpus (14m, 116m)
- Semantic Distance
  - Co-occurence based (G = google)
  - Context-based using web snippets (S)
- All experiments: training on ANEW seed words (cross-validation)

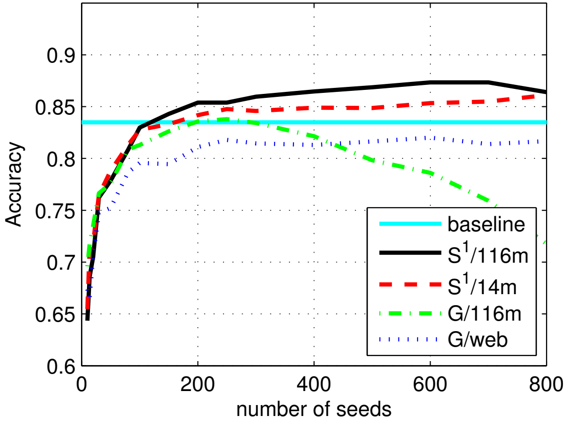
# Word Polarity Detection (ANEW)

2-class word classification accuracy (positive vs negative)



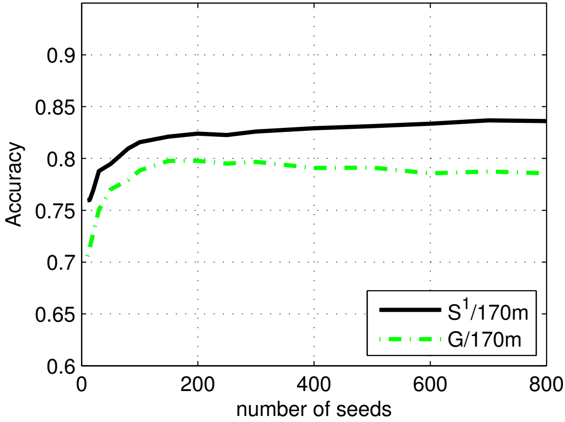
# Word Polarity Detection (GINQ)

2-class word classification accuracy (positive vs negative)



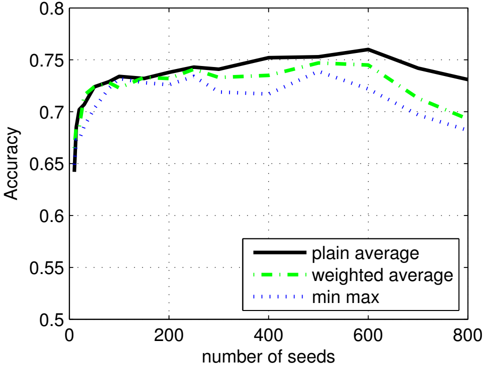
# Word Polarity Detection (BAWLR)

2-class word classification accuracy (positive vs negative)



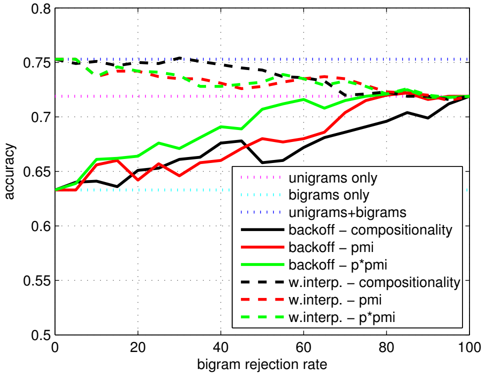
# Sentence Polarity Detection (SemEval 2007)

2-class sentence classification accuracy (positive vs negative), using weighted interpolation



# Sentence Polarity Detection (SemEval 2007)

2-class sentence classification accuracy (positive vs negative), vs bigram rejection threshold



# ChIMP Sentence Frustration/Politeness Detection

- ChIMP Children Utterances corpus
- 15.585 English sentences, Politeness/Frustration/Neutral ratings
- SoA results, binary accuracy P vs 0 / F vs O:
  - 81% / 62.7% [Yildirim et al, '05]
- 10-fold cross-validation
- ANEW used for training/seeds to create word ratings
- ChiMP words added to ANEW with weight $w$, to adapt to the task
- Similarity metric: Google semantic relatedness
- Only content words taken into account

| Politeness: Sentence Classification Accuracy | Fusion scheme | | |
|---|---|---|---|
| | avg | w.avg | max |
| Baseline: P vs O | 0.70 | 0.69 | 0.54 |
| Adapt $w = 1$: P vs O | 0.74 | 0.70 | 0.67 |
| Adapt $w = 2$: P vs O | 0.77 | 0.74 | 0.71 |
| Adapt $w = \infty$: P vs O | **0.84** | 0.82 | 0.75 |
| **Frustration: Sentence Classification Accuracy** | **Fusion scheme** | | |
| | avg | w.avg | max |
| Baseline: F vs O | 0.53 | 0.62 | **0.66** |
| Adapt $w = 1$: F vs O | 0.51 | 0.58 | 0.57 |
| Adapt $w = 2$: F vs O | 0.49 | 0.53 | 0.53 |
| Adapt $w = \infty$: F vs O | 0.52 | 0.52 | 0.52 |

# Summary of Results

- The word-level ratings are very accurate and robust across different corpora
- N-gram sentence-level ratings significantly better than the state-of-the-art, despite the simplistic sentence level fusion model and disregard of syntax/negations
- Adaptation provided good performance on the politeness detection task (linear fusion)
- The baseline model performed best on the frustration detection task (max fusion)

# Conclusions

Proposed a high-performing, robust, general-purpose and scalable algorithm for affective lexicon creation

- Investigated linear and non-linear sentence level fusion schemes, showing good but task-dependent performance
- Investigated domain adaptation with good but task-dependent performance (politeness vs frustration detection task)
- Demonstrated that distributional approach can generalize to n-grams

# Ongoing Work

- Similarity metrics on semantic networks
  - Graph theoretic approaches, e.g., cliques
  - Local and global normalization
- (Non-)compositional models Semantics and Affect:
  - Additional information, modifiers, functionals: syntax, negations, modifiers
  - Fusion of semantic and distributional models
  - Temporal integration of sentence ratings
  - Modeling context and affective reversal
- Cognitive models of semantics and affect
  - Low dimensional semi-metric semantic spaces