A Game – Theoretic Approach to Word Sense Disambiguation

Rocco Tripodi
Marcello Pelillo

Introduction

- **New model** for Word Sense Disambiguation (WSD) formulated in terms of evolutionary game theory
- **Word Sense Disambiguation**: the task of identifying the intended meaning of a word based on the context in which it appears
- “**bank**”
  - Financial sense
  - Naturalistic sense

Background

- Approaches based on

1. Learning models
    - Supervised
    - Unsupervised
    - Semi-supervised
2. Techniques
    - Heuristic
3. Algorithms
    - Graph-based
    - Knowledge-based

✓ Remain on the surface of the word, compromising the coherence of the data to be analyzed

Game Theory

- **Mathematical** approach to study the **interaction** between two or more individuals
  - Benefits and costs depend on the strategies of each other
  - **Players** $I = \{1, .., n\} \rightarrow N$: total number of players
  - **Pure player strategy set** $M_i = \{ m_{i1}, m_{i2}, \dots , m_{i|M_i|} \}$, $|M_i|$: total number of pure strategies of player i
  - **Mixed player strategy vector** $\mathrm{x_i} = (x_{i1}, x_{i2}, \dots, x_{i|M|})$, x_{ih} : probability of player i to choose its $h-th$ pure strategy
    - $\Delta = \{ \mathrm{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, \text{where } \mathrm{x_i} \geq 0 \; for \; i = 0,1, \dots, \mathrm{n} \}$ → probability simplex
    - n : number of pure strategies of player i
  - **Payoff** → value associated with a possible outcome of a game (set of strategies)
  - **Utility** function $u_i : M_1 \times M_2 \times \dots \times M_N \rightarrow \mathbb{R}$
    - strategies → payoffs
    - combination of strategies played by all players
  - **Nash equilibrium** → set of strategy profiles in which each strategy is a **best response** to the strategy of the co-players and no player has the incentive to deviate from their decision (changing strategy ↛ payoff increase)

- **Classical game theory VS Evolutionary game theory**
  - **Static** VS **Dynamic** strategies
  - **Evolutionary game theory** → dynamics of strategy change through repeated games

Proposed System

Evolutionary game theory framework

- **First attempt** in the specific NLP task of WSD

WSD

- Sense - **labelling** task → sense assignment to words
- **Constraint** satisfaction problem

Game-theoretical approach

- **Players** → words (to be disambiguated)
- **Strategies** → senses (evolving population)
- **Payoff** matrices → sense similarity
- **Interactions** → weighted graph
- **Nash equilibrium** → consistent word-sense assignment
- **Selection process**
  - Iterative process of **fitness increase** (candidates / senses with certain features)
  - **Best candidates** (senses with higher fitness) in the population

Proposed System

Consistent final labelling of the data

The solution of the problem is always found

- System convergence to the nearest Nash equilibrium (Nash Theorem 1951)

Most appropriate sense association

- Target word

WSD

- Continuous optimization problem
- Exploitation of contextual information in a dynamic way (evolutionary game theoretic framework)

Versatile approach

- Adaptive to different scenarios and tasks
- Unsupervised / Semi-supervised

Data Modelling

1. Data Geometry
2. Strategy Space
3. Payoff
4. System Dynamics & Sense Classification

Data Geometry

- **List** I of N **words** from the text
  - $I = \{i_1, i_2, \dots, i_N\}$

- **Word similarity** matrix $W_{N\times N}$
  - Pairwise **similarities** among words → players' interactions
  - $w_{kj} = \mathrm{sim}\left(i_k, i_j\right),\ \forall\ k,j = 1,2\ \dots,\ N\ \ and\ \ k \neq j$
  - Weighted adjacency matrix of the graph

- **Similarity** measure: strength of **co-occurrence** between words i, j
  - **Semantically correlated** words
  - 8 **association** measures: Dice coefficient (dice), modified Dice coefficient (mDice), pointwise mutual information (pmi), t-score measure (t-score), z-score measure (z-score), odds ration (odds-r), chi-squared test (chi-s), chi-squared correct (chi-s-c).

- **Proximity** relations with n - neighbours (similarity **augmentation**)
  - Sentence **structure**
  - Size of $\boldsymbol{n}$: fixed expressions / semantic concepts

Data Geometry

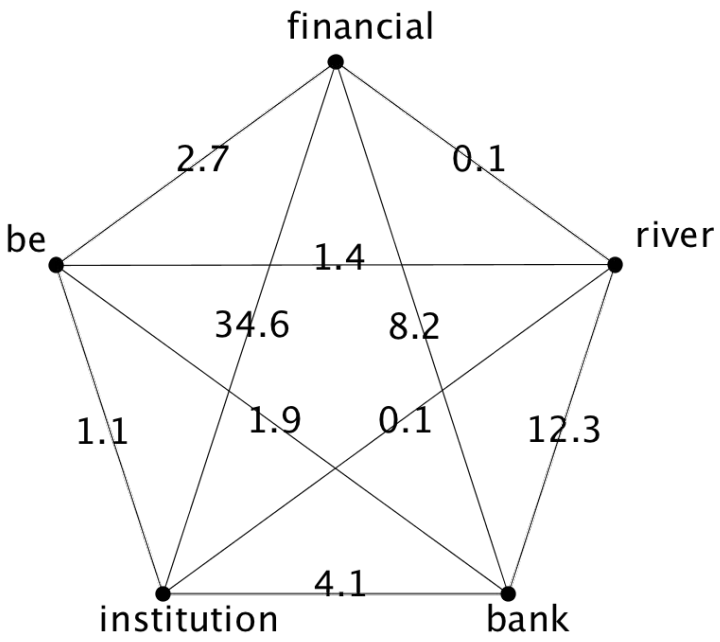


Figure 1. Co-occurrence graph

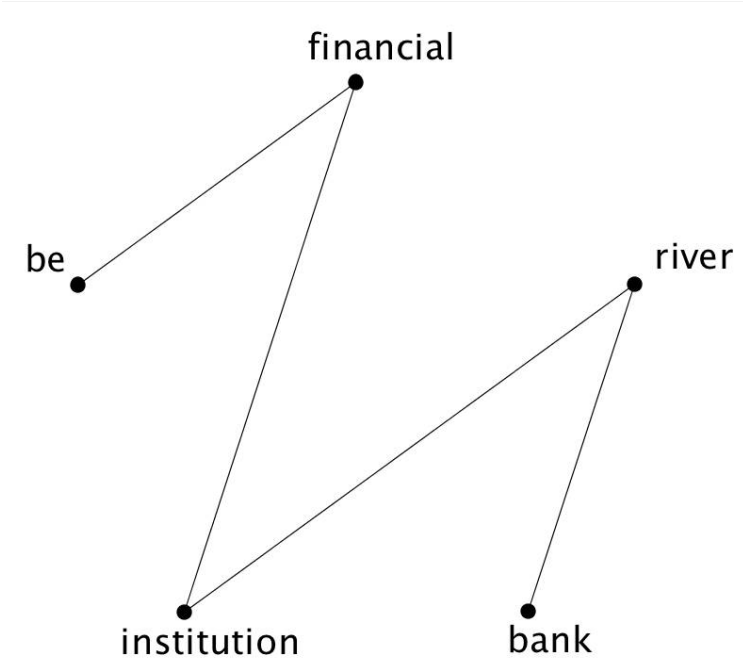


Figure 2. n-gram graph (n=1)

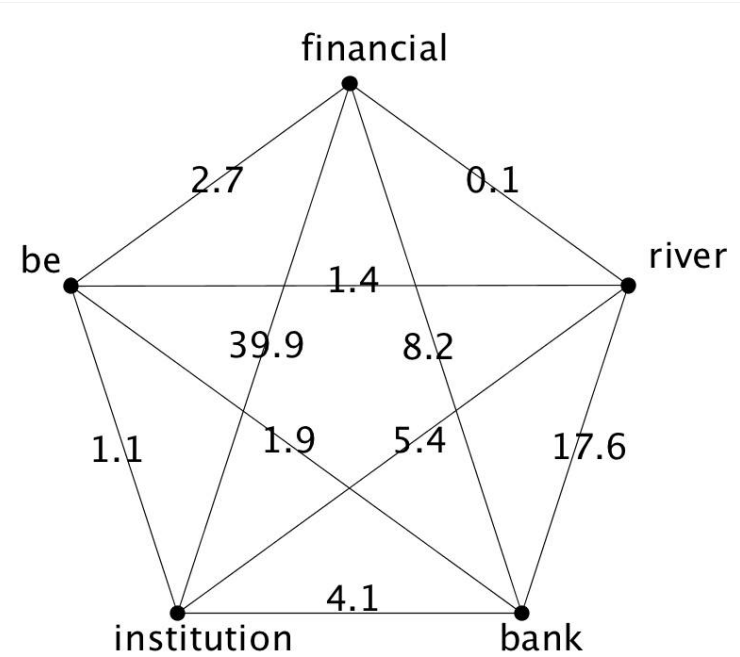


Figure 3. Similarity n-gram graph

Strategy Space

- **Sense inventories** of each word $i \in I$
  - $M_i = \{ m_{i1}, m_{i2}, \ \dots \ , \ m_{ik} \}$, where $k = |M_i|$ the number of **senses** associated with word i
  - WordNet and BabelNet as knowledge base

- **Unique concepts** in sense inventories
  - Game space $C = \{ c_1, c_2, \ \dots, c_k \}$, where $k = |C|$ the total number of unique concepts
  - Assignment of **probability distribution** over the senses in C, for each word $i \in I$

- **Player strategy space**
  - $s_p = \left(s_{p1,} \, s_{p2}, \ \dots \ , \ s_{pn} \right), \ \ n = |C|$
  - s_{pi} : **probability** that player p chooses the $i - th$ sense / concept
  - Graphical representation: regular **polygon** of radius 1, where the distance from the centre to any node - mixed strategy, represents the probability associated with a particular word sense - pure strategy (for $i \in M_i$)

Strategy Space

- **System strategy space** $S_{\mathrm{N}\times|C|}$ initialization
  - s_{ih} : **probability** of player i to choose the $h-th$ pure strategy to play
  - **Uniform** distribution
    - **unsupervised** learning: no prior knowledge
    - $s_{ih} = \begin{cases} |M_i|^{-1}, & if\ sense\ h\ is\ in\ M_i \\ 0, & otherwise \end{cases}$
  - **Geometric** distribution
    - **semi-supervised** learning: exploits information from prior knowledge
    - $pr_{ih} = \begin{cases} p(1-p)^{r_h}, & if\ sense\ h\ is\ in\ M_i \\ 0, & otherwise \end{cases}$
      - ✓ **higher** probability on senses that have a **high frequency** → r_h : sense rank
    - $pr_{ih} = \begin{cases} p(1-p)^{r_o}, & if\ sense\ h\ is\ in\ cluster\ o \\ 0, & otherwise \end{cases}$
      - ✓ **equal** probability to the senses belonging to a determined **cluster** → r_o : cluster rank
    - $s_{ih} = \frac{pr_{ih}}{\sum_{j\in M_i} pr_{ij}} \rightarrow \sum_{h\in C} s_{ih} = 1$

Payoff

- **Sense similarity** matrix $Z_{|C|\times|C|}$
  - Pairwise similarity among senses in $C \rightarrow$ partial payoff matrices of each game
  - $z_{ij} = ssim(c_i,\ c_j),\ \ \forall c_i, c_j \in C,\ i \neq j$
- Partial **payoff** matrix $Z_{|M_i|\times|M_j|}$:
  - Games played between two words i, j
  - Dimension $|M_i| \times |M_j|$, $|M_k|$: cardinality of senses set of word k
- Semantic **similarity**
  - Relations of **likeness** (“is-a”)
  - wup: path depth of senses $c_i,\ c_j$ and msa
    - $ssim_{wup}(c_i,\ c_j) = 2 * \frac{depth(msa)}{depth(c_i)+depth(c_j)}$
    - msa : “most specific ancestor”
  - jcn measure: corpus statistics and structural properties
    - $ssim_{jcn}(c_i,\ c_j) = IC(c_1) + IC(c_2) - 2\,IC(msa)$
    - $IC(c) = -\log(p(c)) = log\,\frac{1}{(p(c))}$
    - Information Content (IC): level of "surprise" of a particular outcome

Payoff

- Semantic **relatedness**
  - Similarity among the **definitions** of two concepts
  - **Wider range** of relations (“is-a-part-of”, “is-the-opposite-of”)
  - Definitions derived from glosses of the synsets in WordNet
  - **Co-occurrence** vector $v_i = (w_{i1},\ w_{i2}\ ,\ \dots\ ,\ w_{in})$, i : concept, w : word gloss occurrences, n : total words
  - **Cosine** similarity
    - $sim(v_i,\ v_j) = \cos\theta = \frac{v_i \cdot v_j}{\|v_i\|\ \|v_j\|}$, i, j : concepts
    - $\|v_i\| = \sqrt{\sum_{j=1}^{n} {w_{ij}}^2}$
    - cosine of the angle θ between the two co-occurrence vectors $v_i,\ v_j$
  - 4 **variations** (construction of vectors)
    - Difference in co-occurrence calculation, corpus use and relation source
    - tf-idf, tf-idf_ext, vec and vec_ext

System Dynamics & Sense Classification

- Computation of the **Nash equilibrium**
  - **Payoff of sense** h (word i)
    - $u_i(e^h, x) = \sum_{j \in N_i} (w_{ij} \, Z_{ij} \, x_j)_h$, where N_i the neighbours (context) of word i
    - Pure strategy h : e^h
    - Similarity with word j : w_{ij}
    - Similarities among senses of i, j : Z_{ij}
    - Sense preference of word j : x_j (s_{jh})
  - Average **payoff of player** ($u_i(x)$)
  - **Replicator Dynamics Equation**
    - $x_i^h(t+1) = x_i^h(t) \, \frac{u_i(e^h, x)}{u_i(x)}$, $\forall \, h \in M_i$ (x_i^h : s_{jh})
    - Player strategy update
    - Depends on performance → measured relatively to the average payoff

System Dynamics & Sense Classification

- **Classification**
  - Mixed strategy $s_{ih} \in S \rightarrow$ each word $i \in I,\ h \in M_i$
  - The strategy with the **highest probability** is chosen
    - $\varphi_i = argmax_{h \in [1,2,\ldots,|C|]}\ s_{ih}$, where $|C|$ the total number of senses
    - Each word $i \in I$ is mapped to exactly one sense $c \in C$
  - Word **fails** to be disambiguated
    - Unable to update its strategy space
    - Strategy space initialized with a uniform distribution
      - Zero entry payoff (no similar senses with co-players)
      - Not connected with other nodes in the graph

System Dynamics & Sense Classification

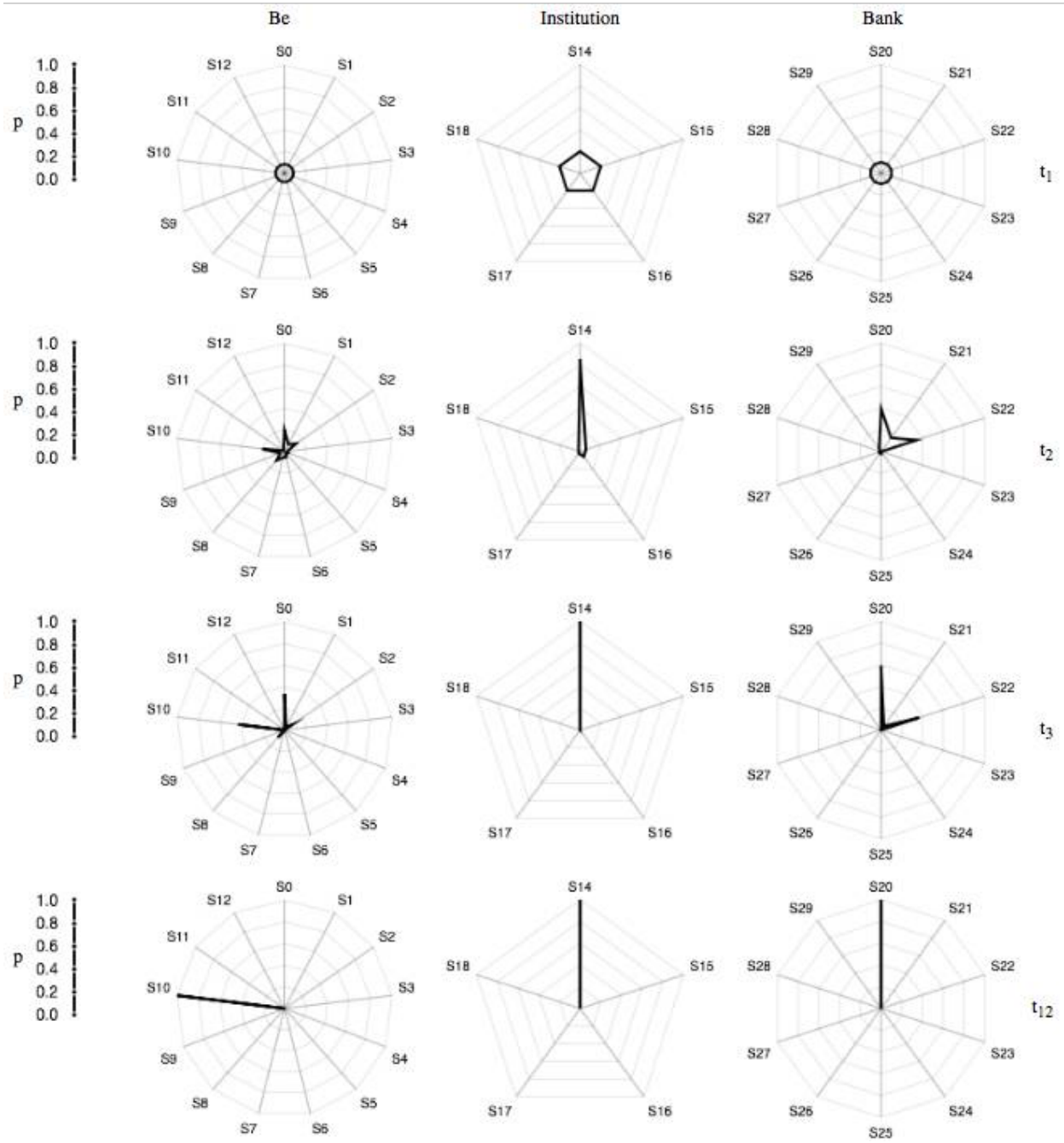


Figure 4. System dynamics for words “be”, “institution” and “bank”

Parameter Tuning

- Two datasets to **evaluate** our algorithm in **different scenarios**
- From each data set
  - 50 different data sets to simulate a situation in which the system must work on texts of different sizes and on different domains
- The **parameters** that will be **tuned**
  - Association, similarity and relatedness measures to weight the similarity among word and senses
  - The n-gram graph to increase the weights of near words (proximity)
  - The p of the geometric distribution used by their semi-supervised system

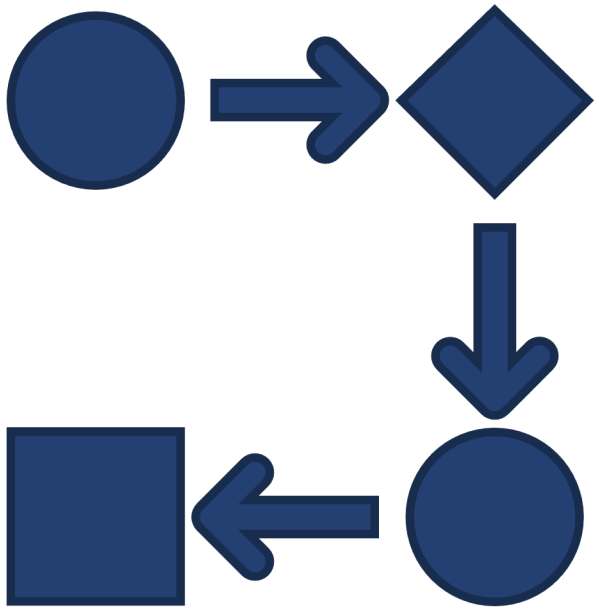
Parameter Tuning - Results

1. Association, Semantic & Relatedness **Measures**
   - The **relatedness** measures perform significantly **better** than the semantic **similarity** measures
   - Particularly **suitable** measures for the algorithm
     - Modified Dice coefficient (**mdice**)
     - Term Frequency – Inverse Document Frequency (**Tf-idf**)

2. $\boldsymbol{n}$ **- gram** Graph
   - Highest results on both data sets using $\boldsymbol{n = 5}$ (n nearest neighbours)

3. **Geometric** Distribution:
   - Best results obtained with $\boldsymbol{p = 0.4}$

Evaluation set-up



- **Results** are provided as balanced **F-score** (F_1) measure
  - F_1: determines the weighted **harmonic mean** of precision and recall
  - $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \times 100$ (%)

- **Evaluation** based on
  - Experiments with **WordNet** as Knowledge base
    - The best performance on nouns on all the data sets
    - Low results on verbs on all data sets
    - The use of prior knowledge is beneficial in general domain data sets
  - Experiments with **BabelNet** as Knowledge base
    - We used BabelNet to collect the sense inventories of each word to be disambiguated and NASARI to obtain the semantic representation of each sense
    - This data set contains highly ambiguous mentions

Comparison to state-of-the-art algorithms

WordNet:

- Our unsupervised system performs better than any other unsupervised algorithm in all datasets
- The performance of our system is more stable on the four datasets
- The comparison with semi-supervised systems shows that our system always performs better than the most frequent sense heuristic when we use information from sense-labeled corpora

BabelNet:

- The performance of our system is close to the results obtained with Babelfy on S13 and substantially higher on KORE50
- It is also difficult to exploit distributional information on this data set because the sentences are short and, in many cases, cryptic.

Comparison to state-of-the-art algorithms

		S7CG	S7CG (N)	S7	S3	S2
unsup.	*Nav10*	–	–	43.1	52.9	–
	PPR_{w2w}	80.1	83.6	41.7	57.9	59.7
	WSD_{games}	**80.4***	**85.5**	**43.3**	**59.1**	**61.2**
semi sup.	*IRST-DDD-00*	–	–	–	58.3	–
	MFS	76.3	77.4	54.7	62.8	65.6*
	MRF-LP	–	–	50.6*	58.6	60.5
	Nav05	**83.2**	84.1	–	60.4	–
	PPR_{w2w}	81.4	82.1	48.6	63.0	62.6
	WSD_{games}	82.8	**85.4**	**56.5**	**64.7***	**66.0**
sup.	*Best*	82.5	82.3*	**59.1**	65.2	**68.6**
	Zhong10	**82.6**	–	58.3	**67.6**	68.2

Figure 5. Experiments with WordNet. Comparison with state-of-the-art algorithms: unsupervised (*unsup.*), semi-supervised (*semi sup.*), and supervised (*sup.*) .

	S13	KORE50
WSD_{games}	**70.8**	**75.7**
Babelfy	69.2	71.5
SUDOKU	66.3	–
MFS	66.5*	–
PPR_{w2w}	60.8	–
KORE	–	63.9*
GETALP	58.3	–

Figure 6. Experiments with BabelNet. Comparison with state-of-the-art algorithms on WSD and entity linking. The results are provided as F1 for S13 and as accuracy for KORE50.

Conclusion

A new method for WSD

- Evolutionary Game Theory
- Similarity measures that perform better
- Continuation of knowledge-based, graph-based approaches

WSD as a constraint-satisfaction problem

- Consistency in the assignment of senses to related words
- Development of contextual coherence on the assignment of senses (characteristic missing in many state-of-the-art systems)

Replicator dynamics equation

- Best labelling assignment

Versatile

- Unsupervised
- Semi-supervised

Competitive compared to state-of-the-art systems

- Considers the influence of each word on the others
- Imposes sense compatibility among each sense before assigning a meaning
- The meaning of a word depends only on words that share a proximity relation and on those that enjoy a high distributional similarity

“ After so much **talking** we got **hungry**, therefore a **toast** is what we need the most ! ”

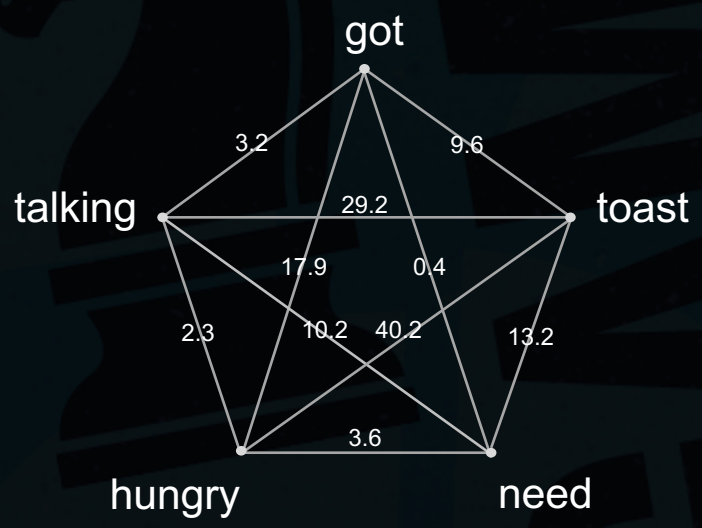


Figure 7. Co-occurrence graph

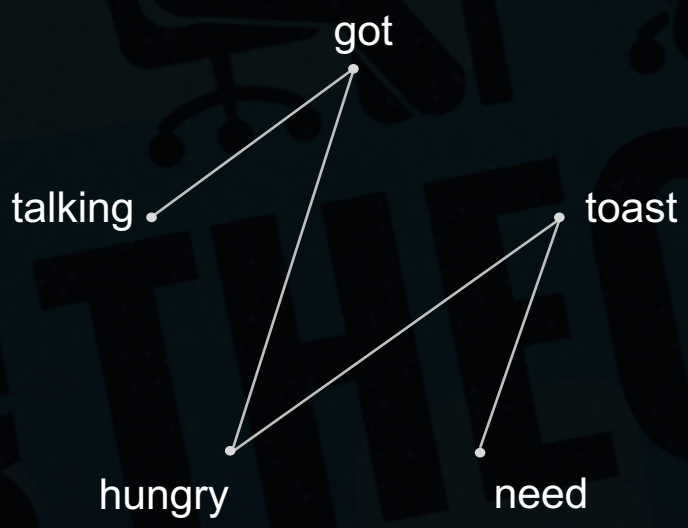
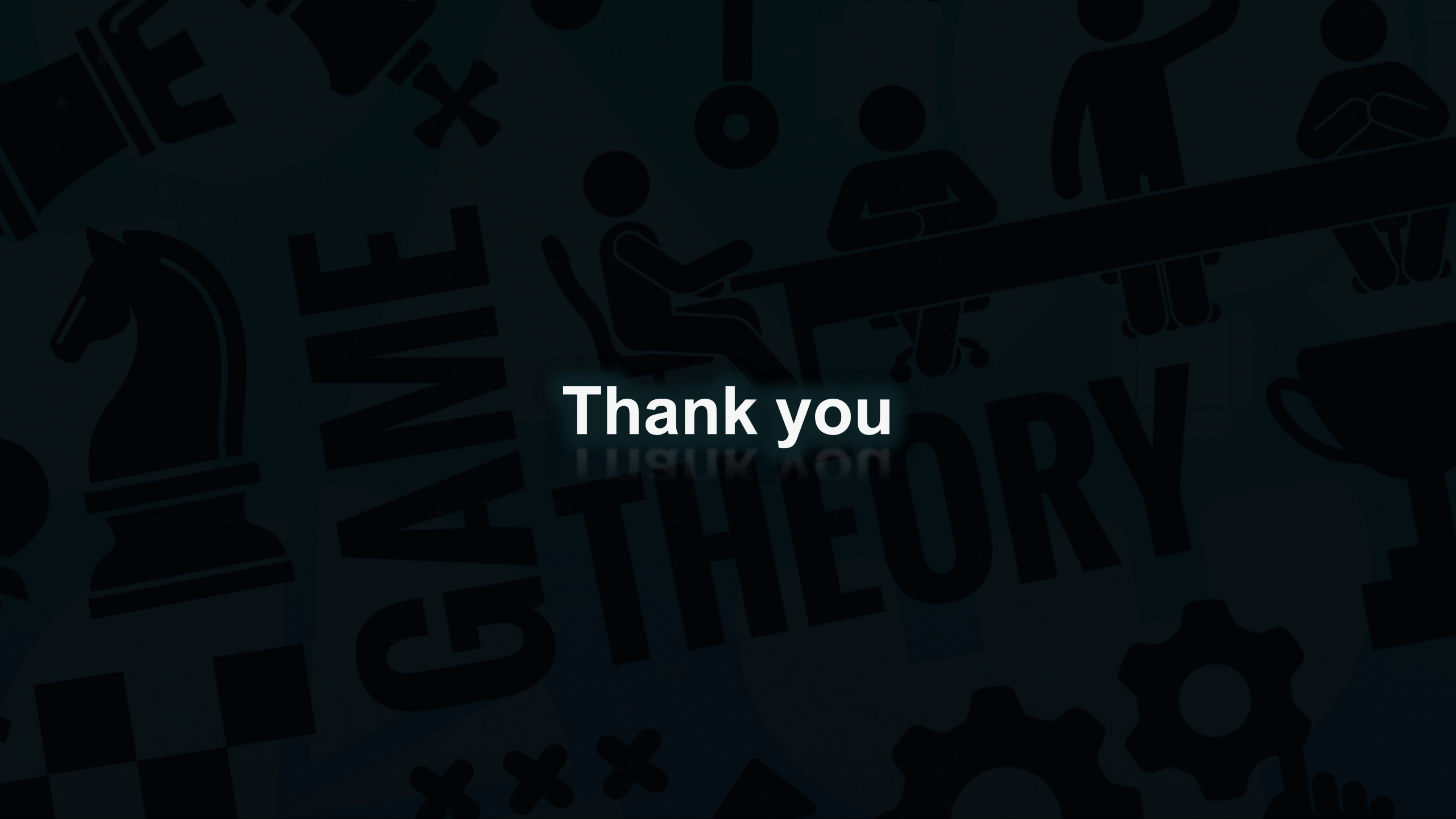


Figure 8. n-gram graph (n=1)

Figure 9. Similarity n-gram graph


Thank you
THEORY

M902 – "Βασικές Μαθηματικές Έννοιες στη Γλωσσική Τεχνολογία"

Kylafi Christina-Theano
Piriasi Juliana

M.Sc. in Language Technology

Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens
&
Institute for Language and Speech Processing,
"Athena" Research Center