A Game-Theoretic Approach to Word Sense Disambiguation

Slides

Slide 1: Paper title, authors, our names, etc

Slide 2: Introduction

1. **a new model for word sense disambiguation formulated in terms of evolutionary game theory**
2. **Word Sense Disambiguation (WSD)**: the task of identifying the intended meaning of a word based on the context in which it appears
3. **Example** → There is a financial institution near the river bank

Spoken text:
Οι συγγραφείς προσεγγίζουν το WSD μέσω του game theory.

Slide 2b: Game Theory

1. A mathematical approach to understand the outcomes of interactions between two or more individuals when benefits and costs of the interactions depend on the strategies of each individual.(+ definitions)
    a. **players** I (= {1, .., n})
    b. **pure strategies** for each player Si (= {s1, ..., sm})
    c. **mixed strategies x :** A mixed strategy set can be defined as a vector x=(x1,...,xm), where m is the number of pure strategies and each component xh denotes the probability that player i chooses its hth pure strategy
    d. **utility** function ui (: S1 × ... × Sn → R) : associates strategies to payoffs

2. **Classical game theory VS Evolutionary game theory**
    a. **Static** strategies (**CGT**) VS **Dynamic** strategies (**EGT**)
    b. **Evolutionary game theory** (EGT)**:**
        i. overcoming some limitations of traditional game theory
        ii. Evolutionary game theory differs from classical game theory in focusing more on the dynamics of strategy change (επανάληψη).

c. **Nash equilibria**: strategy profiles in which each strategy is a best response to the strategy of the co-player and no player has the incentive to deviate from their decision (no way to do better)
d. **Payoff** matrix:

3. WSD problem modelling in game theoretical terms
   a. players -> words
   b. strategies -> senses (evolving population)
   c. payoff matrices -> sense similarity
   d. interactions -> weighted graph
   e. It assumes that there is a **population** of individuals, represented by all the **senses** of the words to be disambiguated, and that there is a **selection process**, which selects the **best candidates** in the population. The selection process is defined as a **sense similarity function**, which **gives a higher score to candidates with specific features**, increasing their fitness to the detriment of the other population members. This process is repeated until the fitness level of the population regularizes and at the end the candidates with higher fitness are selected as solutions of the problem.

- **Game theory** provides predictive power in interactive decision situations. It was intro- duced by Von Neumann and Morgenstern (1944) in order to develop a mathematical framework able to model the essentials of decision-making in interactive situations. In its normal form representation (which is the one we use in this article) it consists of a finite set of players **I** = {1, .., n}, a set of pure strategies for each player Si = {s1, ..., sm}, and a utility function ui : S1 × ... × Sn → R, which associates strategies to payoffs. Each player can adopt a strategy in order to play a game; and the utility function depends on the combination of strategies played at the same time by the players involved in the game, not just on the strategy chosen by a single player.
- **Classical game theory VS Evolutionary game theory**
- A game theoretic framework can be considered as a solid tool in decision-making situations because a fundamental theorem by Nash (1951) states that any normal-form game has at least one mixed Nash equilibrium, which can be used as the solution of the decision problem.
- **Static** strategies (**CGT**) VS **Dynamic** strategies (**EGT**)
- **Evolutionary game theory** (**EGT**) is the application of game theory to evolving populations in biology. It defines a framework of contests, strategies, and analytics into which Darwinian competition can be modelled.
- Natural Selection
- payoffs -> reproductive success
- supervised and knowledge-based. Supervised algorithms learn, from sense-labeled corpora, a computational model of the words of interest. Then, the obtained model is used to classify new instances of the same words. Knowledge-based algorithms perform the disambiguation task by using an existing lexical knowledge base, which usually is structured as a semantic network. Then, these approaches use graph algorithms to disambiguate the

words of interest, based on the relations that these words' senses have in the network (Pilehvar and Navigli 2014).

- The WSD problem can be formulated in game-theoretic terms modeling:
-
- •the **players** of the games as the words to be disambiguated.
- •the **strategies** of the games as the senses of each word.
- •the **payoff matrices** of each game as a sense similarity function.
- •the **interactions** among the players as a weighted graph.
- **Nash equilibria** correspond to consistent word-sense assignments!
-
- •**Word-level similarities**: proportional to strength of co-occurrence between words
- **Sense-level similarities**: computed using WordNet / BabelNet ontologies
-

Slide 3 : Background

- Multiple NLP applications can **benefit** from the disambiguation of ambiguous words, as a preliminary process; otherwise they remain on the surface of the word, compromising the coherence of the data to be analyzed
- But to the best of our knowledge**,** our method is **the first attempt to use it in a specific NLP task**
- **supervised :** model learns from sense-labeled corpora and then is obtained to classify new instances
- **knowledge-based :** using an existing lexical knowledge base, use graph algorithms to disambiguate the words based on the relations that these words' senses have in the network
- **unsupervised :** learns patterns from untagged data
- **semi-supervised :** small amount of labelled & large amount of unlabelled data
- **heuristics :** practical (not necessarily optimal) solution to a problem
- **graph-based :** model the relations among words and senses in a text with graphs, representing words and senses as nodes and the relations among them as edges

Slide 4 : Proposed system

1. Our approach is defined in terms of **evolutionary game theory**.
2. WSD task
    a. **sense-labeling task** (sense assignment to word)
    b. **constraint satisfaction problem**
3. WSD problem modelling in game theoretical terms
    - **players** → words
    - **strategies** → senses (evolving population)
    - **payoff** matrices → sense similarity

- **interactions** → weighted graph
- **Nash equilibria** correspond to consistent word-sense assignments
- **Word-level similarities**: proportional to strength of co-occurrence between words
- **Sense-level similarities**: computed using WordNet / BabelNet ontologies
- It assumes that there is a **population** of individuals, represented by all the **senses** of the words to be disambiguated, and that there is a **selection process**, which selects the **best candidates** in the population. The selection process is defined as a **sense similarity function**, which **gives a higher score to candidates with specific features**, increasing their fitness to the detriment of the other population members. This process is repeated until the fitness level of the population regularizes and at the end the candidates with higher fitness are selected as solutions of the problem.

4. This approach ensures that the final labeling of the data is consistent and that the solution of the problem is always found. In fact, our system **always converges to the nearest Nash equilibrium** from which the dynamics have been started . **TODO - always converges**
5. This approach gives us the possibility not only to exploit the **contextual** information of a word but also to find **the most appropriate sense association for the target word and the words in its context.**
6. Within a game theoretic framework we are able to cast the WSD problem as a continuous optimization problem, **exploiting contextual information <u>in a dynamic way</u>**
7.
8. **Versatile approach:** method is adaptive to different scenarios and to different tasks, and it is possible to use it as **unsupervised** or **semi-supervised** (new semi-supervised version of the approach, which can exploit the evidence from sense tagged corpora or the most frequent sense heuristic and does not require labeled nodes to propagate the labeling information.).

Slides 5++ : Data modelling & Math

(p.43)

- Brief info on dataset (WordNet, corpus used, etc)
- Περιγραφή του βήματος - μαθηματική ανάλυση (σημειογραφία / τύποι / πίνακες / τεχνική εφαρμογή / γραφήματα-περιγραφή / φυσική σημασία)
- Παράλληλη αναφορά στο παράδειγμα με το river bank (φλεξ με prisoner’s dilemma - game theory έννοιες)

Slide 6: Step 2 - 3 - **Geometry of data (players’ interactions)**

1. Compute from I the word **similarity matrix W** in which are stored the pairwise similarities among each word with the others and represents the players’ interactions
2. **similarity matrix** W_{NxN} (where N the total number of words retrieved)
    a. $w_{kj} = sim(i_k , i_j), \ \forall k,j \in I : k \neq j$

3. Data geometry: weighted adjacency matrix of the weighted graph (Figure 4a)
4. Similarity measure: strength of co-occurrence between words i, j
    a. reflects the relations of semantically correlated words
5. Target words not in reference corpus -> query expansion (alternative lexicalisations of lemma using WordNet)
6. Proximity relations with n-neighbours (similarity augmentation - step 3)
    a. reflects the sentence structure
    b. small n: fixed expressions
    c. large n: semantic concepts
    d. n-gram structure (graph representation Figure 4b)
    e. similarity n-gram graph (Figure 4c)
7. No disconnected nodes
8. 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).
9.

- graph construction (5.1.1)
- co-occurrence graph of example (Figure 4a)
    - graph: geometry of the data
    - nodes: words
    - weighted edges: m-dice similarity(association measure) between words i,j
    - WordNet

Slide 7: Step 4 - 5 - **Strategy space**
1. collect **sense inventories** of each word M using a knowledge base (using WordNet and BabelNet as knowledge bases)
2. player (word) strategy space
    a. represented as a regular polygon of radius 1, where the distance from the center to any vertex (mixed strategy) represents the probability associated with a particular word sense (pure strategy) (Figure 5)
3. create the **list** C = (1, . . . , c) of all the unique concepts in the sense inventories (space of game)
4. assign for each word in I a **probability distribution** over the senses in C (probability distribution over the possible strategies)
5. Strategy space S (matrix)
    a. Each component sih denotes the probability that the player chooses to play its hth pure strategy among all the strategies in its strategy profile

Slide 9: Step 6 - **Payoff**
1. compute the sense similarity matrix Z among each pair of senses in C, which is then used to compute the partial payoff matrices of each game
    a. $z_{ij} = ssim(i_i\ ,\ i_j),\ \forall i,j \in I : i \neq j$
2. partial payoff matrix: single games played between two players i and j, Z_ij -> dim mxn, m, n senses of words i, j.

3. semantic similarity
    a. wup: only the path length among two concepts (τύπος?!)
        i. $ssim(s_i, s_j)$ = 2 ∗ depth(msa)/(depth(s_i) + depth(s_j))
        ii. $ssim(s_i, s_j)$ = IC(s_i) + IC(s_j) − 2IC(msa), where IC(c) = "information content - surprisal - of concept c" - IC(c)= -log_2(p(c))- and msa= "most specific ancestor node"
    b. jcn measure: corpus statistics and structural properties (τύπος!?)
4. semantic relatedness
    a. similarity among the definitions of two concepts
    b. definitions derived from glosses of the synsets in WordNet
    c. co-occurrence vector vi = (w1,i,w2,i...wn,i) (for each concept i, where w represents the number of times word w occurs in the gloss - n total words)
    d. cosine similarity
        i. $ssim(s_i, s_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}$, where $\|v_i\| = \sum_{j=1}^{n} w_{ji}$ (cosine of the angle between the two co-occurrence vectors
    e. 4 variations (way the gloss vectors are constructed)
        i. difference in co-occurrence calculation, the corpus use, and relation source
        ii. tf-idf, tf-idf_ext, vec, and vec_ext.

Slide 10: Steps 7-8 - **System Dynamics & Sense Classification**

1.
2. assign to each word *i* ∈ *I* a strategy *s* ∈ *C* unique -> argmax (φ - equation 7)
3.

- Apply the replicator dynamics equation in order to compute the Nash equilibria of the games
- If each player has chosen a strategy—an action plan choosing his own actions based on what it has seen happen so far in the game—and no player can increase his own expected payoff by changing his strategy while the other players keep theirs unchanged, then the current set of strategy choices constitutes a **Nash equilibrium**.
- **Nash equilibria** correspond to **consistent word-sense assignments**!
- There are several reasons for the **prominent role of replicator dynamics:** Firstly,(1) is relatively simple and mathematically well understood (Hofbauer and Sigmund2003). It is equivalent to the famous Lotka–Volterra model in population ecology(Hofbauer1981). Additionally, there are beautiful connections between replicator dynamics and the concepts of classical game theory (Fudenberg and Tirole1991;We i bu l l1995). For example, **if strategy i is dominated**, meaning that **there is another strategy which is always better**, then **replicator dynamics will lead to the extinction of i** (through computations - **equation 6**)

Slide 11: Step 8

- assign to each word *i* ∈ *I* a strategy *s* ∈ *C*

Slides 12-13: Parameter Tuning & Experiments

- We **tested** our approach on different data sets from WSD and **entity-linking tasks** in order to find the **similarity measures that perform better**, and **evaluated** our **approach** against **unsupervised, semi-supervised, and supervised state-of-the-art systems**.
- (WordNet+BabelNet + algorithm comparison + graphs)
- Τελικές παράμετροι - με βάση προηγούμενα πειράματα κλπ κλπ
    - mdice (association measure - weighted edges of graph W - word similarity)
    - tf-idf (payoffs)
    - n=5 (n-gram)
    - p=0.4 (semi-supervised learning)
- F-score (F1) - μαθηματικός τυπος (αρμονικός μεσος)
    - **precision:** Precision represents the proportion of items – in this case, entities – that the system returns which are accurately correct. It rewards careful selection, and punishes over-zealous systems that return too many results: to achieve high precision, one should discard anything that might not be correct.False positives – spurious entities – decrease precision.
    - **recall:** Recall indicates how much of all items that should have been found, were found. This metric rewards comprehensiveness: to achieve high recall, it is better to include entities that one is uncertain about. False negatives – missed entities – lead to low recall. It balances out precision.
    - **Precision:** το ποσοστό των οντοτήτων που κατηγοριοποιήθηκαν (αποσαφηνίστηκαν) σωστά προς το σύνολο οντοτήτων που κατηγοριοποιήθηκαν από το μοντέλο
    - **Recall:** το ποσοστό των σωστά κατηγοριοποιημένων οντοτήτων (από το σύνολο όλων)
    - In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been).
    - Precision is the estimated probability that a document randomly selected from the pool of retrieved documents is relevant.
    - Recall is the estimated probability that a document randomly selected from the pool of relevant documents is retrieved.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

-
-
- The results of this evaluation show that **our method performs well and can be considered as a valid alternative to current models**.
- **WordNet** as knowledge-base
- **BabelNet** as knowledge-base

Parameter tuning

•Two data sets (S10 and S15) to tune the parameters: from specific vs different domains→ all the content words of the second data set have to be disambiguated

A. association and semantic measures to weight the similarity among word and senses

B. the n-gram graph to increase the weights of near words

C. the *p* of the geometric distribution used by their semi-supervised system

A. Association and Semantic Measures:

•The performance of the system is highly influenced by the combination of measures used [Tables 2-3]

•The relatedness measures perform significantly better than the semantic similarity measures→ the representations provided by the relatedness measures show a block structure on the main diagonal of the matrix, which is exactly what is required for a similarity measure.

•(The use of the tf-idf weighting schema seems to be able to reduce the noise in the data representation)

•If we observe the performances of the association measures we notice that on average the best measures are dice, mdice, chi-s-c, and also odds-r on S15

B. N-gram Graph:

•To test the influence, we selected the association and relatedness measures with the highest results with increasing values of n this approach is always beneficial for S15, BUT on S10 not always beneficial.

•Sometimes an improvement on S10: we notice that the pair of measures with highest results on both data sets is tfidf-mdice with n = 5. This also confirms our earlier experiments in which we saw that these two measures are particularly suited for our algorithm.

C. Geometric Distribution:

→ we want to exploit information from sense-labeled corpora

•we used the pair of measures and the value of n detected with the previous experiments, increasing values of p, in the interval [0.05, 0.95].

•the performance of the semi-supervised system on S15 is always better than that obtained with the unsupervised system (p = 0) **BUT** performance on S10 is always lower than that obtained with the unsupervised system (because target words of S10 belong to a specific semantic domain)

•From the plot we can see that on S15 the highest results are obtained with values of p ranging from 0.4 to 0.7 and for the evaluation of our model we decided to use p = 0.4 as parameter for the geometric distribution, because with this value we obtained the highest result

Error Analysis:

•The main problems are related to the semantic measures

•The adverbs and adjectives are not disambiguated with these measures because of the lack of payoffs (low semantic content) but also in verbs with rich semantic content (e.g. generate, prevent, obtain)

•Use of n-gram graph to increase the weights among neighboring words noticed that when used with the relatedness measures it leads to the disambiguation of all the target words and with n ≥ 1 we have precision = recall

•Using the local information given by the n-gram graph, allows us to balance the influence of words in the text

•Another aspect to consider is whether the polysemy of the words influences the results of the system the use of information from sense-labeled corpora is particularly useful when the polysemy of the words is particularly high

Slide 14: Key Findings
(all evaluation results, bullet points)

Evaluation set-up

•Algorithm evaluated with three fine-grained data sets and one coarse-grained data set using as the knowledge base, **WordNet**

• Evaluated our approach on two data sets using as the knowledge base, **BabelNet**

•Recall that for all the next experiments we used mdice to weight the graph W, tfidf to compute the payoffs, n = 5 for the n-gram graph, and p = 0.4 in the case of semi-supervised learning

Experiments Using WordNet as Knowledge Base:

•The best performance of our system is obtained on nouns on all the data sets.This is in line with state-of-the-art systems because in general the nouns have lower polysemy and higher inter-annotator agreement

- ❖ our method is particularly suited for nouns (the disambiguation of nouns benefits from a wide context and local collocations)

•We obtained low results on verbs on all data sets common problem not only for supervised and unsupervised WSD systems

•This series of experiments confirms that the use of prior knowledge is beneficial in general domain data sets and that when it is used, the system performs better than the most common-sense heuristic computed on the same corpus.

Comparison to State-of-the-Art Algorithms:

•Compared our method with supervised, unsupervised, and semi-supervised systems on four data sets

•Our unsupervised system performs better than any other unsupervised algorithm in all data sets (Table 5)

•(In S3 and S2, the difference is more substantial compared with both unsupervised systems)

•The performance of our system is more stable on the four data sets, showing a constant improvement on the state-of-the-art

•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

•We note <u>strange behavior on S7CG</u> (++) but on the other three data sets we note a substantial improvement in the performances of our system, <u>with stable results higher than state-of-the-art systems</u>
•we note that the results of our semi-supervised system on the fine-grained data sets are close to the performance of state-of-the-art supervised systems

•We also note that the <u>performance of our system on the nouns of the S7CG data set is higher than the results of the supervised systems [Table 4]</u>

Experiments with BabelNet:

•We used BabelNet to collect the sense inventories of each word to be disambiguated, the mdice measure to weight the graph W, and NASARI to obtain the semantic representation of each sense (The only difference with the experiments presented in Section 6.2.1 is that we used BabelNet as knowledge base and NASARI as resource to collect the sense representations instead of WordNet)

•This data set <u>contains highly ambiguous mentions</u> that are difficult to capture without the use of a large and well-organized knowledge base. In fact, the mentions are not explicit and require the use of common knowledge to identify their intended meaning

Comparison to State-of-the-Art Algorithms:

•The performance of our system is close to the results obtained with Babelfy on S13 and substantially higher on KORE50 because <u>with our approach</u> <u>it is necessary to respect the textual coherence</u> (which is required when a sentence contains a high level of ambiguity, such as those proposed by KORE50)

•T<u>he good performance of our approach is also due to the good semantic representations</u> provided by NASARI able to exploit a richer source of information

•(It is also difficult to exploit distributional information on this data set because the sentences are short and, in many cases, cryptic the recall on this data set is well below the precision: 55.5%)

•+++(Able to correctly disambiguate the same entities in sentences where there is more contextual information)

Slide 15: Conclusion & Strengths

1. **The** exploitation of the **<u>contextual</u>** <u>information</u> of a word along with the fact that we also find **the most appropriate sense association for the target word and the words in its context, is the most important contribution of our work,** which distinguishes it from existing WSD algorithms. In fact, in some cases using only contextual information without the imposition of

constraints can lead to inconsistencies in the assignment of senses to related words. (Μπήκε στο proposed)

2. This idea can preserve the **textual coherence**—a characteristic that **is missing in many state-of-the-art systems**. In particular, it is missing in systems in which the words are disambiguated independently. On the contrary, **our approach disambiguates all the words in a text concurrently**, using an underlying structure of interconnected links, which **models the interdependence between the words**. In so doing, we model the idea that **the meaning for any word depends at least implicitly on the combined meaning of all the interacting words**

3. Furthermore**, no supervision is required** and **the system can adapt easily to different contextual domains,** which is exactly what is required for a WSD algorithm

Slide 16: Smart closing
- n-gram / co-occurrence graph με **αμφίσημες** λέξεις / βάρη, κλπ

Notes

- Φυσική σημασία **equation 6** : η πιθανότητα (t+1) να επιλεγεί η έννοια h από τη λέξη i εξαρτάται από την πιθανότητα (t) και το άθροισμα των προτιμήσεων των υπόλοιπων λέξεων ως προς αυτή την έννοια (utility - payoff) διαιρούμενη από το μέσο όρο (average) των προτιμήσεων όλων των λέξεων για όλες τις έννοιες.
- **Relative fitness:** https://en.wikipedia.org/wiki/Fitness_(biology): Relative fitnesses only indicate the change in prevalence of different genotypes relative to each other, and so only their values relative to each other are important; relative fitnesses can be any nonnegative number, including 0.
- A **fitness function** is a particular type of objective function that is used to summarise, as a single figure of merit, how close a given design solution is to achieving the set aims. Fitness functions are used in genetic programming and genetic algorithms to guide simulations towards optimal design solutions.
- The **expected utility hypothesis** is a popular concept in economics, game theory and decision theory that serves as a reference guide for judging decisions involving uncertainty.[1] The theory recommends which option a rational individual should choose in a complex situation, based on his tolerance for risk and personal preferences. The expected utility of an agent's risky decision is the mathematical expectation of his utility from different outcomes given their probabilities
-

Definitions

- **Altruism**: When one organism reduces its own fitness to benefit the fitness of another organism.
- **Evolutionary stable strategy**: A behavioral strategy (phenotype) if adopted by all individuals in a population that cannot be replaced or invaded by a different strategy through natural selection.
- **Game theory**: A mathematical approach to understanding the outcomes of interactions between two or more individuals when benefits and costs of the interactions depend on the strategies of each individual.
- **Inclusive fitness**: The fitness of a gene as measured by the fitness of the individual possessing the gene and the fitness of the individual's relatives bearing the same gene, identical by descent.
- **Mutualism**: A relationship between two individuals from different species that benefits each individual involved in the interaction.
- **Phenotype**: The physical, physiological, behavioral and other traits expressed by an individual.

Cheat Sheet

1. The WSD problem can be formulated in game-theoretic terms modeling:

•the players of the games as the words to be disambiguated.
•the strategies of the games as the senses of each word.
•the payoff matrices of each game as a sense similarity function.
•the interactions among the players as a weighted graph.
Nash equilibria correspond to consistent word-sense assignments!
•Word-level similarities: proportional to strength of co-occurrence between words
Sense-level similarities: computed using WordNet / BabelNet ontologies

Game theory offers a principled and viable solution to context-aware pattern recognition problems, based on the idea of dynamical competition among hypotheses driven by payoff functions.
Distinguishing features:
•No restriction imposed on similarity/payoff function (unlike, e.g., spectral methods)
•Shifts the emphasis from *optima* of objective functions to *equilibria* of dynamical systems.
On-going work:
•Learning payoff functions from data (Pelillo and Refice, 1994)
•Combining Hume-Nash machines with deep neural networks
•Applying them to computer vision problems such as scene parsing, object recognition, video analysis

Links

> precision & recall

- **lecture8a.pptx** ← !!!! (έχει ωραίο θέμα - δες μήπως το "δανειστούμε" χιχι) - τελικά ίσως και όχι χαχα
- https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient
- http://repfiles.kallipos.gr/html_books/246/Ch2.html
- https://en.wikipedia.org/wiki/F-score
-