An Integrated Approach to Automatic Synonym Detection in Turkish Corpus

Dr. Tuğba YILDIZ
Assist. Prof. Dr. Savaş YILDIRIM
Assoc. Prof. Dr. Banu DİRİ

İSTANBUL BİLGİ UNIVERSITY
YILDIZ TECHNICAL UNIVERSITY
Department of Computer Engineering

September 17, 2014

Semantic Relation

- Semantic relations are underlying relations betwen two concepts expressed by words or phrases (Moldovan 2004)
- Contradictory views about number of semantic relations
- Uncountability (Murphy 2003)
  - 13 classes (Rosario 2001)
  - 17 classes (Stephens 2001)
  - 5 classes at the top and 30 classes at the bottom (Nastase 2003)
  - 35 classes (Moldovan 2004)
  - 7 classes (SemEval 2007 - Task4)
  - etc.

Semantic Relation

- Five main families of relations:
  - Hyponym/Hypernym (Class inclusion)
  - Meronym/Holonym (Part-Whole)
  - Synonym (Similars)
  - Antonym (Contrast)
  - Case

Semantic Relation

- Five main families of relations:
  - Hyponym/Hypernym (Class inclusion)
  - Meronym/Holonym (Part-Whole)
  - Synonym (Similars)
  - Antonym (Contrast)
  - Case

Semantic Relation

- In this study:
  - Synonym (Similars)
    - Synonyms are words with identical or similar meanings
    - "car is synonymous with automobile"
  - Hyponym/Hypernym (Class inclusion)
    - A relation of inclusion (IS-A / kind of)
    - "A dog is an animal"
    - dog: hyponym / animal: hypernym
  - Meronym/Holonym (Part-Whole)
    - A relationship between terms that respect to the significant parts of a whole
    - "the eye is part of the face"
    - eye: part / face: whole

Related Works

- Methods are mostly based on distributional similarity.
- Distributional hypothesis which adopts that **"semantically similar words share similar contexts"** (Harris 1954).
- However insufficient for synonymy!
  - covers near-synonyms (Eg. top-5 for orange is: yellow, lemon, peach, pink, lime) (Lin et. al. 2003)
  - does not distinguish between synonyms and other relations
- Different strategies:
  - integrating two independent approaches such as distributional similarity and pattern-based approaches
  - utilizing external features such as dependency relations.

Related Works

- In Turkish, recent studies on synonym relations are based on dictionary definition TDK[1] and Wiktionary.
- Defined rules and phrasal patterns that occur in the dictionary definitions with.
- **Objective of this study:** to determine synonymy in a Turkish Corpus
- The main contributions of our work:
  - its corpus-driven characteristics
  - relies on both dependency and semantic relations.

[1] TDK:Turkish Language Association

Language Resources

- BOUN Web corpus with 500M tokens (Sak et.al. 2008)
  - four sub-corpora:
    - Three of them named NewsCor are from three major Turkish news portals
    - GenCor is a general sampling of web pages in the Turkish Language
  - morphological parser based on two-level morphology
  - an averaged perceptron-based morphological disambiguator
- Preprocessing
  - surface+root+POS-tag+[and all other markers]
    Turkish: araba
    English: car
    "arabanın+araba+noun+a3sg+pnon+gen"

Methodology

- No particular LSPs to extract synonymy
- Our main assumption:
  - Synonym pairs mostly show similar dependency and semantic characteristics in corpus.
  - They share the same meronym/holonym relations, same particular list of governing verbs, adjective modification profile and so on.
  - Examples:
    - Automobile **is a** vehicle
    - Car **is a** vehicle
    - Wheel **is part of** automobile
    - Wheel **is part of** car

Methodology

- **Model:** determine if a given word pair is synonym or not.
- **Data:** Manually and randomly selected 200 synonym pairs(SYN) and 200 non-synonym pairs(NONSYN) to build a training data set.
  - Non-synonym pairs are especially selected from associated (relevant) pairs such as tree-leaf, student-school, computer-game, etc.
- **Features from Corpus:** For each synonym pair, 15 different features.
  - Co-occurrence (1)
  - Semantic relations based on LSPs (4)
  - Dependency relations based on syntactic patterns (10)
- **Target Class:** SYN and NONSYN

Features from Corpus

- **Feature 1- Co-occurrence:** The first feature is the co-occurrence of word pairs within a broad context
- **Features 2/3- Meronym/Holonym:** A big matrix in which rows depict Whole candidates, columns depict Part candidates
- Example:

Table: 1. Whole X Part Matrix

Whole/Part	**Part1**	**Part2**	**...**	**Partm**
Whole1	val.	val.	...	val.
Whole2	val.	val.	...	val.
...	val.	val.	...	val.
Wholen	val.	val.	...	val.

- Cells represent the possibility of that corresponding whole and part are in meronymy relation

Features from Corpus

Table: 2. General Patterns (GP), Dictionary-based Patterns(TDK-P) and Bootstrapped Patterns(BP)

GP	**TDK-P**	**BP**
NPx part of NPy	Group-of (whole|group|all) (set|flock|union)	NPy-gen NPx-pos
NPx member of NPy	Member-of (class|member|team) (from the family of Y)	NPy-nom NPx-pos
NPy constituted of NPx	Amount-of (amount|measure|unit)	NPy-Gen (N-ADJ)+ NPx-Pos
NPy made of NPx	Has/Have (Y has I(H))	NPy of one-of NPx
NPy consist of NPx	Consist-of	NPx whose NPy
NPy has/have NPx	Made-of	NPxs with NPy
NPy with NPx		

Features from Corpus

- To measure the similarity of meronymy/holonymy profile of two given words, cosine function is applied on two rows/columns indexed by two given words

Table: 3. Example of Whole X Part Matrix

Whole/Part	Wheel	...	Car	....	Wiper	Automobile	...
Car	X	...	...	...	X	...	...
School	...	...	...	...	...	...	...
Person	...	...	...	...	...	...	...
Automobile	X	...	...	...	X	...	...
Parking	...	...	X	...	...	X	...

Features from Corpus

- **Features 4/5 - Hyponym/Hypernym:** Another matrix is built for hypernymy/hyponymy by applying LSPs
- Same procedure is carried out as Meronym/Holonym.
  - "NPs gibi CLASS" (CLASS such as NPs)
  - "NPs ve diğer CLASS" (NPs and other CLASS)
  - "CLASS lArdAn NPs" (NPs from CLASS)
  - "NPs ve benzeri CLASS" (NPs and similar CLASS)

Features from Corpus

- **Features 6–15 - Dependency Relations:** are obtained by syntactic patterns
- The more they are modified by same adjectives, the more likely they are synonym.
- Examples:
- Fast Car (+) / Fast Automobile (+) / Handsome Car (X) / Handsome Automobile (X)
- 36 different patterns were extracted, 8 were eliminated because of the poor results.
- Then we grouped them according to their syntactic structures.

Features from Corpus

Table: 4. Dependency Features

Features	Dependency relation	# of Patterns	Examples
G1	direct object of verb	13	I drive a car araba sürüyorum
G2	subject of verb	3	waiting car bekleyen araba
G3	direct object/subject of verb	3	- -
G4	modified by adjective+(with/without)	2	car with gasoline benzinli araba
G5	modified by inf	1	swimming pool yüzme havuzu
G6	modified by noun	1	toy car oyuncak araba
G7	modified by adjective	1	red car kırmızı araba
G8	modified by acronym locations	1	the cars in ABD ABD'deki arabalar
G9	modified by proper noun locations	1	the cars in Istanbul Istanbul'daki arabalar
G10	modified by locations	2	the car at parking lot otoparktaki araba

Results and Evaluation

- All features explained before are computed for all pairs/instances
- All computed scores of the pairs are kept as a training set
- Taking all features into account and applying machine learning algorithms
- The most suitable algorithm, logistic regression

Results and Evaluation

- The first aim is to find out which feature is the most informative for detecting synonymy and contributes most to the overall success of the model.

Table: 5. F-Measure of Semantic Relations (SRs) Features

	co-occurrence	hyponym	hypernym	meronym	holonym
F-Measure	62.5	60.5	60	68.7	73.7

- The semantic features are notably better than dependency features.
- Among semantic relations, the most powerful attributes are meronymy and holonymy features.
- The possible reason: the sufficient number of cases matched by lexico-syntactic patterns.

Results and Evaluation

- Among dependency relations, G1, G4 and G7 have better performance

Table: 6. F-measure of Dependency Relations Features

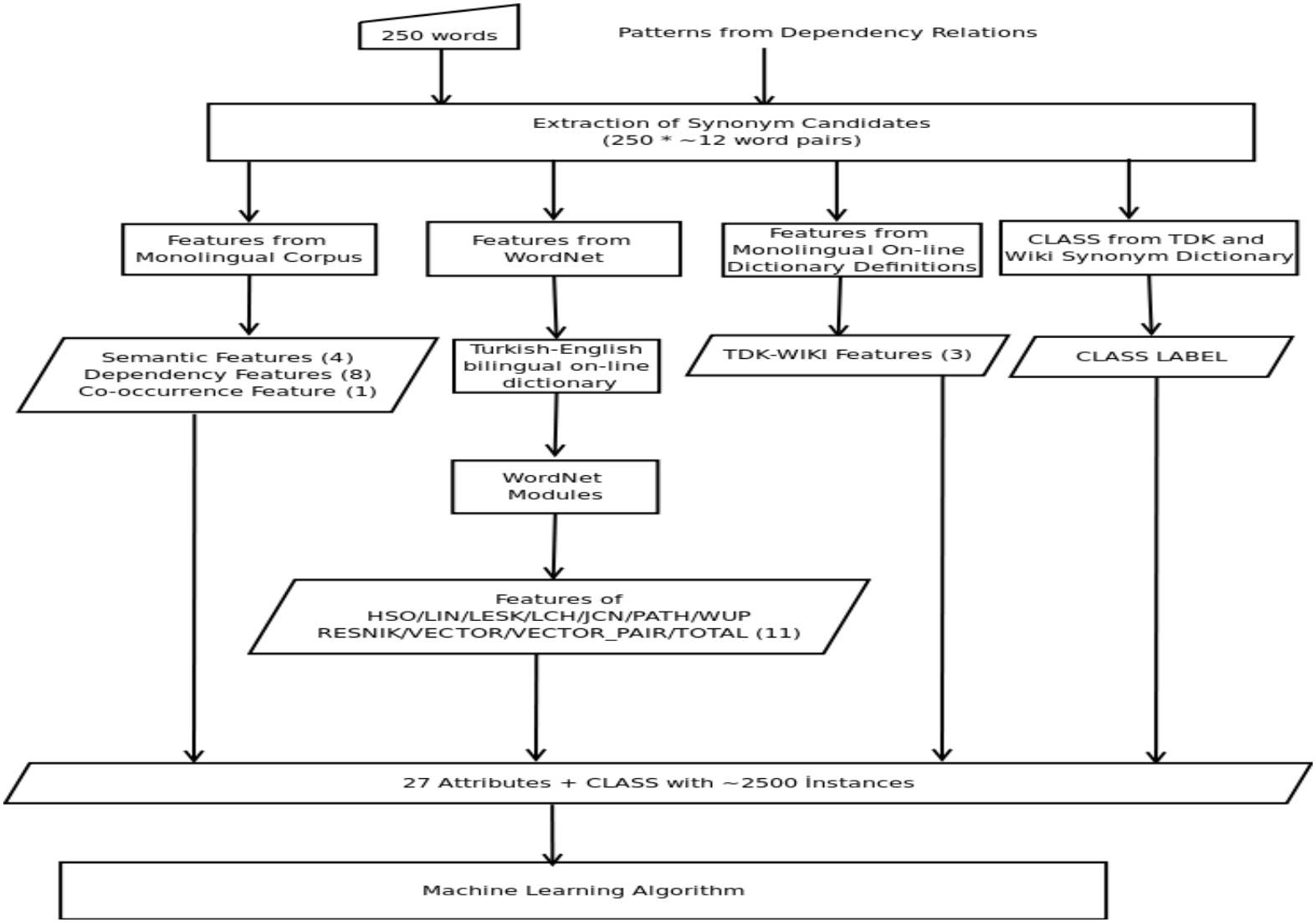
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
F-Measure	**64.7**	58	60.5	**65**	61.6	58.8	**63**	49.4	48.3	62.6

- The poorest groups, G8 and G9, have low production capacity
- The successful features are linearly dependent on target class.
- On the aggregated data where all useful features are considered, the performance of logistic regression is F-measure of 80.3%.
- The achieved score is better than the individual performance of each feature.

Done

- In our paper: "Dictionary definitions, WordNet, and other useful resources could be used and evaluated in future work."
- In our review: "WordNet seems a good resource helping improve it even more."

Done


250 words
Patterns from Dependency Relations
Extraction of Synonym Candidates
(250 * ~12 word pairs)
Features from Monolingual Corpus
Features from WordNet
Features from Monolingual On-line Dictionary Definitions
CLASS from TDK and Wiki Synonym Dictionary
Semantic Features (4)
Dependency Features (8)
Co-occurrence Feature (1)
Turkish-English bilingual on-line dictionary
TDK-WIKI Features (3)
CLASS LABEL
WordNet Modules
Features of
HSO/LIN/LESK/LCH/JCN/PATH/WUP
RESNIK/VECTOR/VECTOR_PAIR/TOTAL (11)
27 Attributes + CLASS with ~2500 İnstances
Machine Learning Algorithm

Conclusion

- Semantic relation is one of the problem in applications in NLP.
- Automatic acquisition of semantic relation has become more popular in recent years.
- It is obviously impossible to discover synonym pairs by applying LSP.
- Thanks to semantic and syntactic relations.
- The most powerful semantic relations: Meronym and Holonym.
- The most powerful syntactic relations: G4 modified by adjective (with/without).
- Aggregation of useful features gives promising result.

Thank You For Listening...