Word **S**ense **I**nduction using methods typical for association rule mining

Phd thesis

Marek Kozłowski

Warsaw, 05.06.2013

POLITECHNIKA WARSZAWSKA

Preface

The times of exponential growth of data

- Mainly unstructured data
- Unlikely to be analyzed by humans
- Currently used approaches are based on lexico-syntactic analysis of text → words occurrences

Two main flaws of the currently used approach are:

- Inability to identify documents using different wordings
- Lack of context-awareness what leads to retrieval documents which are not pertinent to the user needs

Knowledge of an actual meaning of a polysemous word can significantly improve retrieving more relevant documents or extracting relevant information from texts.

Ambiguity

Ambiguity around us

- More than 73% of words in English are polysemous
- Average number of meanings per word is approximately 6
- Apple can be used as a software company name, personal computers, fruit, or river.
- Plant can be used to mean a botanical life form or a industrial building.
- Bass may refer respectively to low-frequency tones and a type of fish.

Sample:

Suppose Robin and Joe are talking, and Joe states, “The bank on the left is solid, but the one on the right is crumbling.” What are Robin and Joe talking about? Are they on Wall Street looking at the offices of two financial institutions, or are they floating down the Mississippi River looking for a place to land their canoe?

Foundations

Word sense theories

- There are many approaches to define sense (referential theory, mentalist, behaviourist and **use theory**)
- „...the meaning of a word is its use in the language“ (Wittgenstein)
- Distributional hipothesis introduced by Zellig Harris, which can be summarized in a few words: „a word is characterized by the company it keeps“
- **Main assumption used by context-based wsd methods: Semantically similar words tend to occur in similar contexts.**

Sample:

wine: beer, white wine, red wine, Chardonnay, champagne, fruit, food, coffee, juice, Cabernet, cognac, vinegar, Pinot noir, milk, vodka,…

Basic concepts

Word Sense Discovery – ambiguous concept

- Word sense disambiguation -
  - process of meaning identification for words in context
  - is the ability to identify the meaning of words in context in a computational manner
  - is an AI-complete problem, that is a problem whose difficulty is equivalent to solving central problems of artificial intelligence.

- Word sense induction -
  - subtask of unsupervised WSD
  - task of automatically identifying senses of words in texts, without the need for handcrafted resources or manually annotated data.

Applications

Possible applications of WSI

- **Information retrieval** e.g: new type of Web Search Engines Or Local Domain Oriented Searchers
- **Information Extraction** e.g: acronym expansion, people disambiguation
- **Question Anserwing** - the main strategy for question answering is to find documents that have the right content even if the same words are not used.
- **Machine Translation** - context to decide about sense of translated word is necessary
- **Lexicographers supporting tool**
- **Support tools for building ontologies** - create relations between concepts and theirs wordings representatives (lexemes)

Methodology

Context

- Set of surrounding words (size is determined by granularity)
- For example:
    - *is the pomaceous fruit of the* ***apple*** *tree, species Malus domestica in the rose family*

Significant context

- Closed frequent set with support greater than minimal threshold (i.e. 2 documents)
- For example:
    - [apple, inc, computer, macintosh]

Sense Frame

- Computational representation of possible sense
- Contains tree multi-level structure, where main context is a root, and sub contexts are nodes

Sense

- Clustered sense frames

Methodology

Main Algorithm – steps done to find senses for a query

Input: set of documents (wikipedia, or another corpora), which is indexed by Lucene

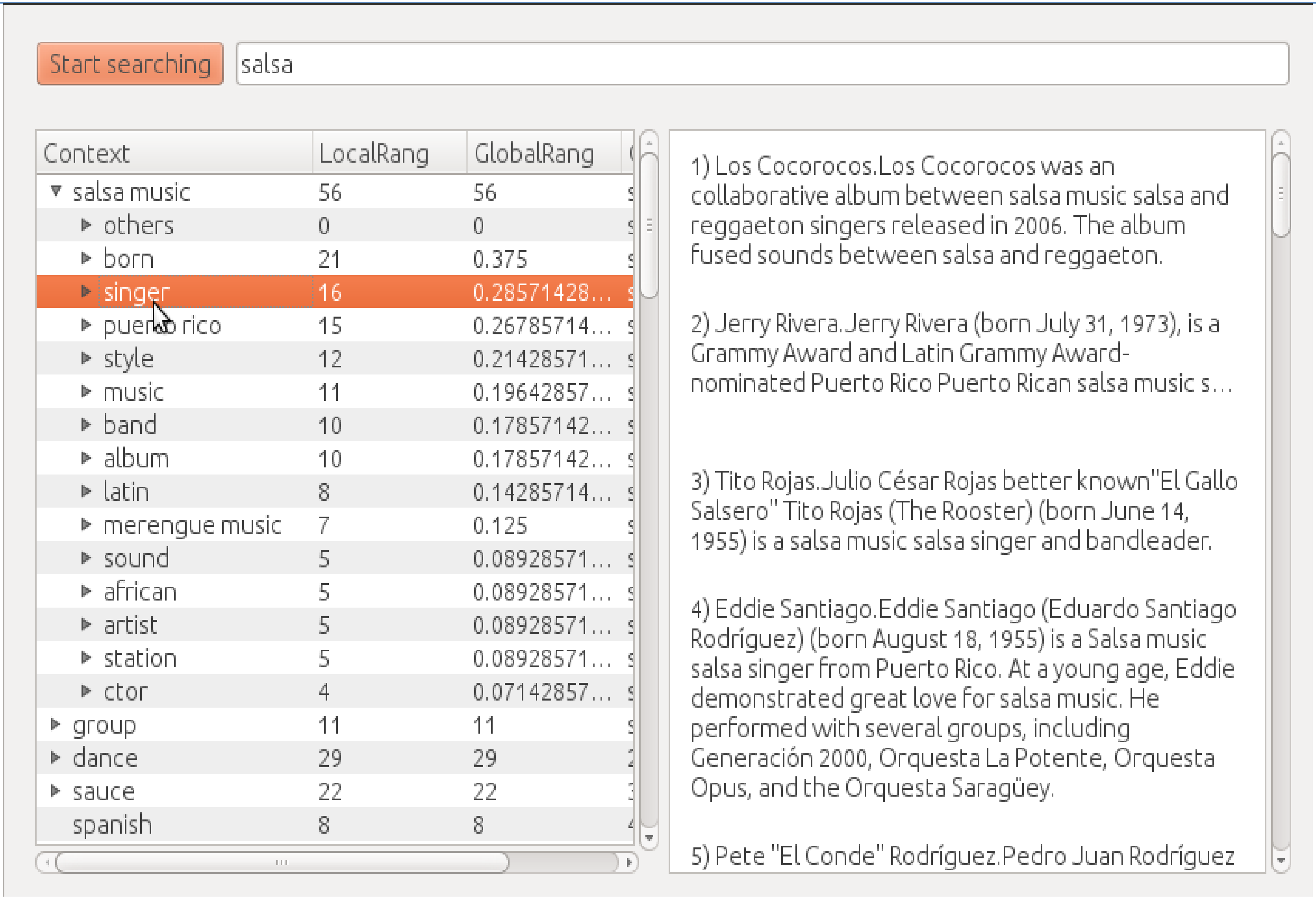
- Using full-text search there are found all paragraphs of documents which contains a given query
- Simple context build stage → only alphanumeric noun-phrases and proper names surrounding the given term are persisted as a **contexts**
- Relevant context build stage → closed frequent sets mining on contexts is taken to build set of **significant contexts**
- Sense frames are constructed from significant contexts (subset-superset relations within a range of context)
- Sense are generated from clustered sense frames

Output for end-user is a set of **senses,** with multi-level hierarchies and matching them documents.

The SnS method

- **The SenseSearcher (SnS)** is a word sense induction algorithm based on closed frequent sets and multi-level sense representation. SnS performed better than methods using vector space modelling. Induced senses by SnS characterize better readability (are more intuitive), also they are hierarchical, what gives them flexible granularity.
- **Key features of SnS are:**
  - ability to find infrequent, dominated senses;
  - number of likely senses determined by content of corpora, there is no fixed threshold determining constant number of retrieved senses;
  - multi-level hierarchies of senses (describing subsenses)

Sample evaluation



Context	LocalRang	GlobalRang	(
▼ salsa music	56	56	s
▶ others	0	0	s
▶ born	21	0.375	s
▶ singer	16	0.28571428…	s
▶ puerto rico	15	0.26785714…	s
▶ style	12	0.21428571…	s
▶ music	11	0.19642857…	s
▶ band	10	0.17857142…	s
▶ album	10	0.17857142…	s
▶ latin	8	0.14285714…	s
▶ merengue music	7	0.125	s
▶ sound	5	0.08928571…	s
▶ african	5	0.08928571…	s
▶ artist	5	0.08928571…	s
▶ station	5	0.08928571…	s
▶ ctor	4	0.07142857…	s
▶ group	11	11	s
▶ dance	29	29	2
▶ sauce	22	22	3
spanish	8	8	4

1) Los Cocorocos.Los Cocorocos was an collaborative album between salsa music salsa and reggaeton singers released in 2006. The album fused sounds between salsa and reggaeton.

2) Jerry Rivera.Jerry Rivera (born July 31, 1973), is a Grammy Award and Latin Grammy Award-nominated Puerto Rico Puerto Rican salsa music s…

3) Tito Rojas.Julio César Rojas better known"El Gallo Salsero" Tito Rojas (The Rooster) (born June 14, 1955) is a salsa music salsa singer and bandleader.

4) Eddie Santiago.Eddie Santiago (Eduardo Santiago Rodríguez) (born August 18, 1955) is a Salsa music salsa singer from Puerto Rico. At a young age, Eddie demonstrated great love for salsa music. He performed with several groups, including Generación 2000, Orquesta La Potente, Orquesta Opus, and the Orquesta Saragüey.

5) Pete "El Conde" Rodríguez.Pedro Juan Rodríguez

Comparison with others

Algorithms which can be used alternatively:

- LSA – Latent Semantic Analysis
  - Co-occurring terms are mapped to the same dimensions, not co-occurring terms are mapped to different dimensions
  - Lower number of dimensions leads to generalizations over the simple frequency data.
- LDA – Latent Dirichlet Allocation
  - probabilistic model of text generation
  - models each document using a mixture over K topics, which are in turn characterized as distributions over words
- Carrot2
  - descriptive documents clustering
  - organize small collections of documents into thematic categories.

Comparison with LDA

- LDA is a probabilistic model of text generation, that is a three-level hierarchical Bayesian model, in which item of a collection is modeled as a finite mixture over an underlying set of topics.
- **Example output:**
  - Group 1 (state, inc, name, dessert, city, computer, student, school, county, year)
  - Group 2 (name, inc, series, family, computer, model, fruit, tree, producer, december)
  - Group 3 (development, airport, game, newton, wine, valley, computer, list, family, center)
  - Group 4 (state, inc, name, dessert, city, computer, student, school, county, year)

Comparison with LSI

- LSI models the meaning of words and documents by projecting them into a vector space of reduced dimensionality, which is built up by applying singular value decomposition
- **Example output:**
  - Group 1 (computer, inc, macintosh, system, software, product, model, mac)
  - Group 2 (country, state, population, censu, river, davies, valley, town, place)
  - Group 3 (-album, -band, -rock, -record, -music, -pie, -tree, population, censu)
  - Group 4 (fruit, cultivar, -album, -pathogenic, -band, -rock, -viru. -family, -plant)

The SnS output

- **SenseSearcher (SnS)** is a word sense induction algorithm based on closed frequent sets and multi-level sense representation.
- **Example output:**
  - Sense 1 (cultivar → fruit, biology, flavor, eating, cider)
  - Sense 2 (american → born, new jersey, company, united states)
  - Sense 3 (apple ii, computer → apple ii series, model)
  - Sense 4 (plant → pathogenic, family)
  - Sense 5 (album → {})
  - Sense 6 (dessert → {})
  - ...

Open problems

Algorithm enhancements

- Effective preprocessing (do as much as You can before)
- Performance
  - Scalability related to significant context generation stage
  - Scalability related to sense frames generation stage
- Clustering of sense frames (verify other methods)
- External knowledge resources (wikipedia > wordnet)
- Database of senses
- ...