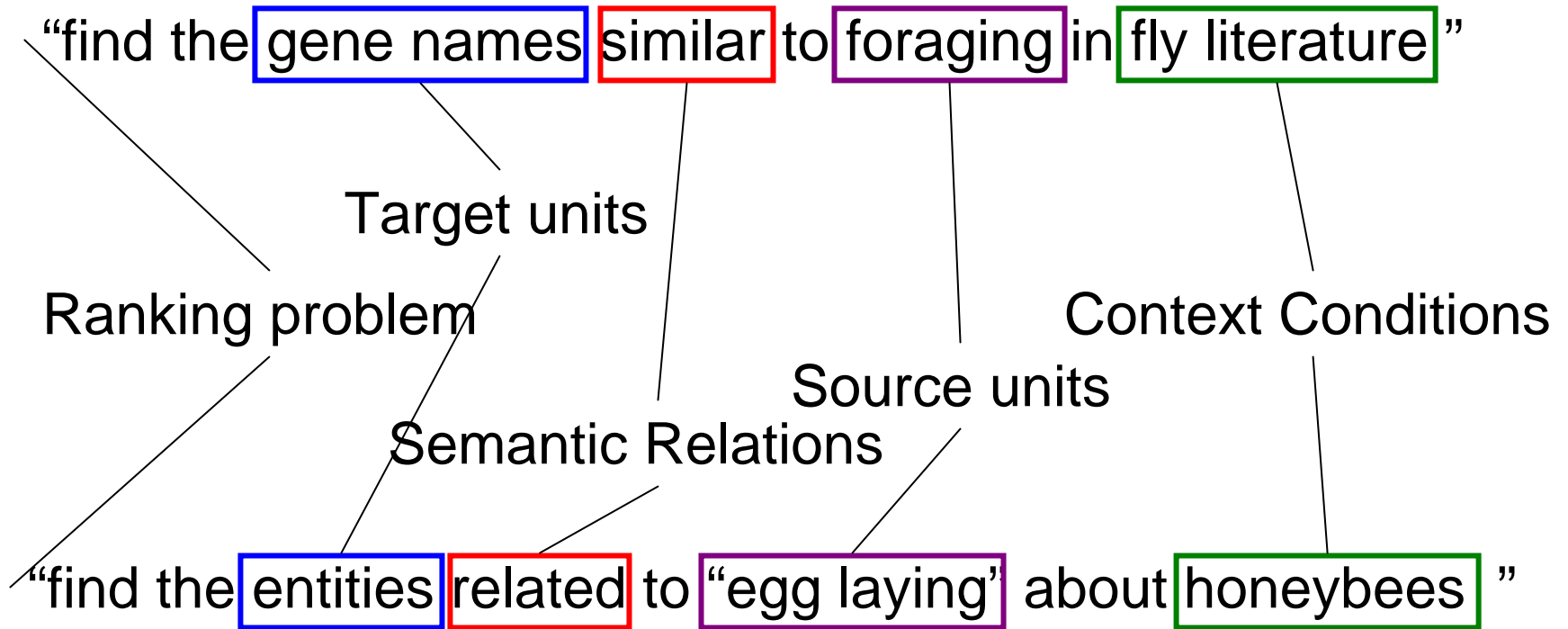


Semantic Processing with Context Analysis

Qiaozhu Mei

2006.10.4

Motivating Examples



Semantic Processing

“find the **gene names** **similar** to **foraging** in **fly literature**”

“find the **sentences** **similar** to **aspect 1** about **gene A**”

“find the **themes** **related** to **“nursing”** in **bee literature**”

“find the **terms** **related** to **“queen”** in **document A**”

.....

These problems are similar in form
→ can we model them in a general way?

What We can Do with Semantic Processing

- User select:
 - Source Unit Type; Target Unit type;
- User specify:
 - Source Unit; Context Conditions
- End-User Applications:
 - Retrieval
 - Categorization/Taxonomy/Encoding
 - Annotation
 - Summarization
 - Synonym Extraction

Context Analysis

“You shall know a word by the company it keeps.”

- Firth 1957

We report the molecular characterization of the spineless (**ss**) gene of [Drosophila](#), and present evidence that it plays a central role in defining the distal regions of both the [antenna](#) and [leg.](#) **ss** encodes the closest known [homolog](#) of the [mammalian dioxin receptor](#), a transcription factor of the [bHLH-PAS](#) family.

-- [biosis:199800271084](#)

Replaceable: Semantically similar

Co-occurring: Semantically Related

Originally reported by Bridges in 1914, **spineless** plays a central role in defining the distal regions of both the [antenna](#) and [leg.](#) **spineless** encodes the closest known [homolog](#) of the [mammalian dioxin receptor](#), a transcription factor of the [bHLH-PAS](#) family. -- <http://www.sdbonline.org/fly/dbzhnsky/spinles1.htm>

Two Basic Types of Semantic Relations

- Semantically related
 - Co-occurring in the same context
 - E.g., spineless → antenna, leg
- Semantically similar
 - Replaceable in the same context
 - E.g., spineless → ss
- Other types of semantic relations are in between.

Dimensions of the Semantic Processing

- Units:
 - Source, target
 - words, phrases, entities, concepts, group of words (clusters), sentences, themes, documents, etc.

Dimensions of the Semantic Processing (II)

- Contexts:
 - Natural contexts: sentences, documents, etc
 - Local contexts: 2-grams; n-grams; window size = k ;
 - Conditional contexts:

Context conditional on "Drosophila"

Local Context:
 $w = 3$

We report the molecular characterization [of the spineless (*ss*) gene] of *Drosophila* and present evidence that it plays a central role in defining the distal regions of both the antenna and leg. *ss* encodes the closest known homolog of the mammalian dioxin receptor, a transcription factor of the bHLH-PAS family.

Natural Context:
sentences

Local Context:
 $w = 5$

Dimensions of the Semantic Processing (III)

- Context Representation:
 - Sets; Vectors; Graphs; Language Models, etc

Dimensions of the Semantic Processing (IV)

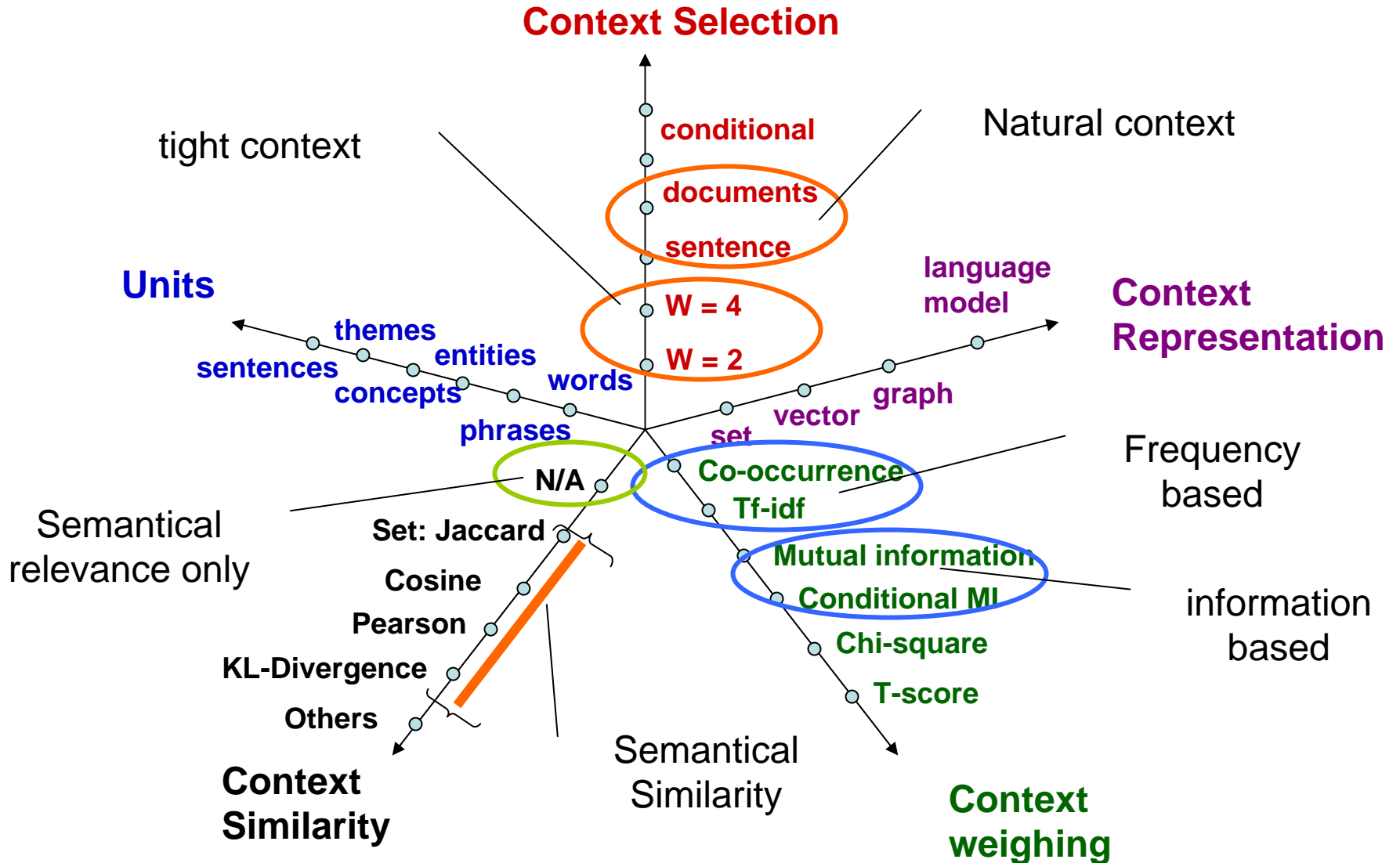
- Context unit weighting:
 - frequency based: co-occurrence, tf-idf, conditional probability, etc
 - MI based: mutual information, pointwise mutual information; conditional mutual information, etc.
 - Hypothesis testing based: Chi-square test, t-test, etc.

Can be used to measure semantic relevance

Dimensions of the Semantic Processing (Cont.)

- Semantic Relations:
 - Semantically related;
 - Semantically similar;
- Similarity functions:
 - Jaccard distance; Cosine; Pearson coefficient; KL divergence; etc.
- Semantic Problems:
 - Ranking
 - Classification
 - Clustering

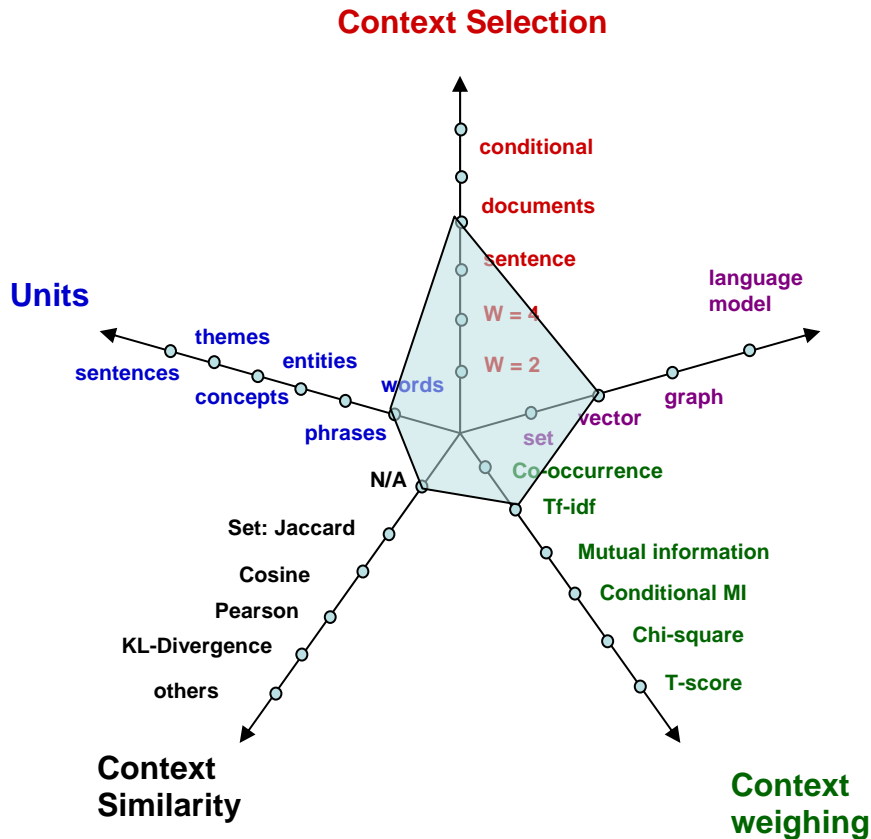
The Space of Semantic Processing



A General Procedure of Semantic Process Problems

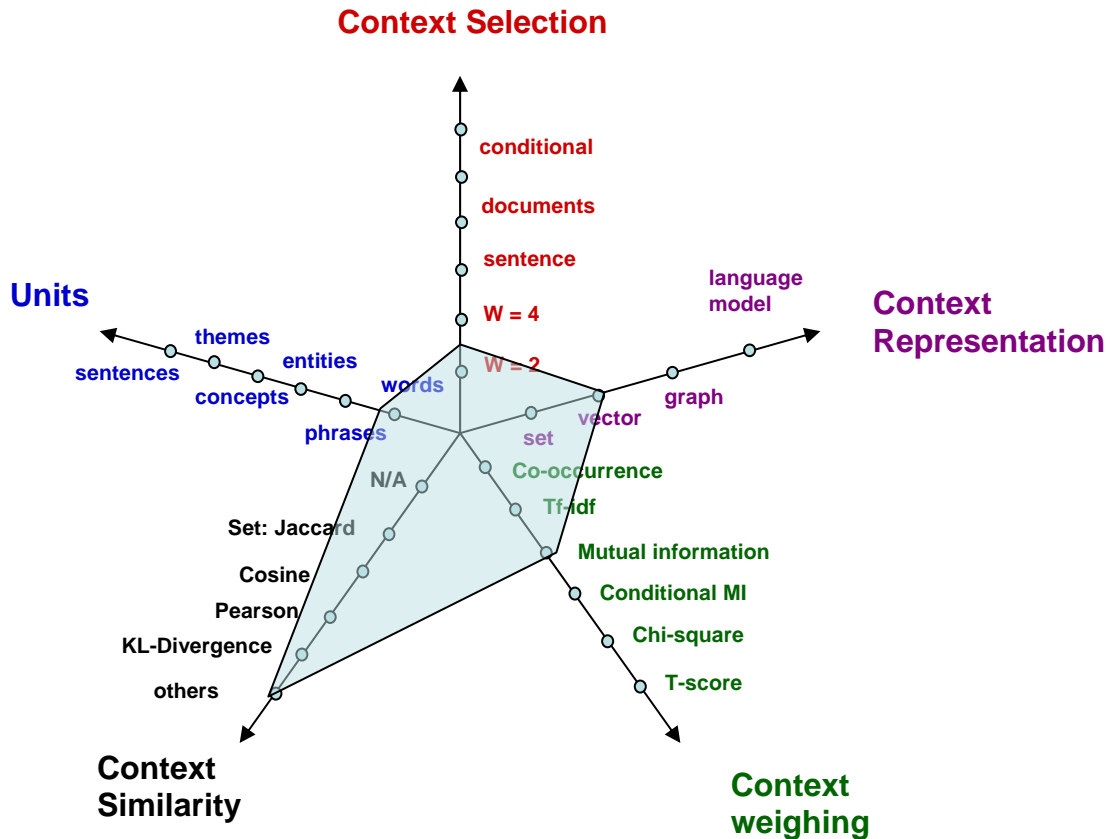
- Input:
 - What are the source units; what are the target units
- Formalization:
 - A ranking problem? Classification problem? Clustering problem?
- Solution (*a ranking problem for example*):
 - Select an appropriate context;
 - Select a representation of the context
 - Select a weighting for the units in the context
 - Select a similarity measure to compare contexts
 - Ranking based on the weight or similarity

Example: Query Expansion



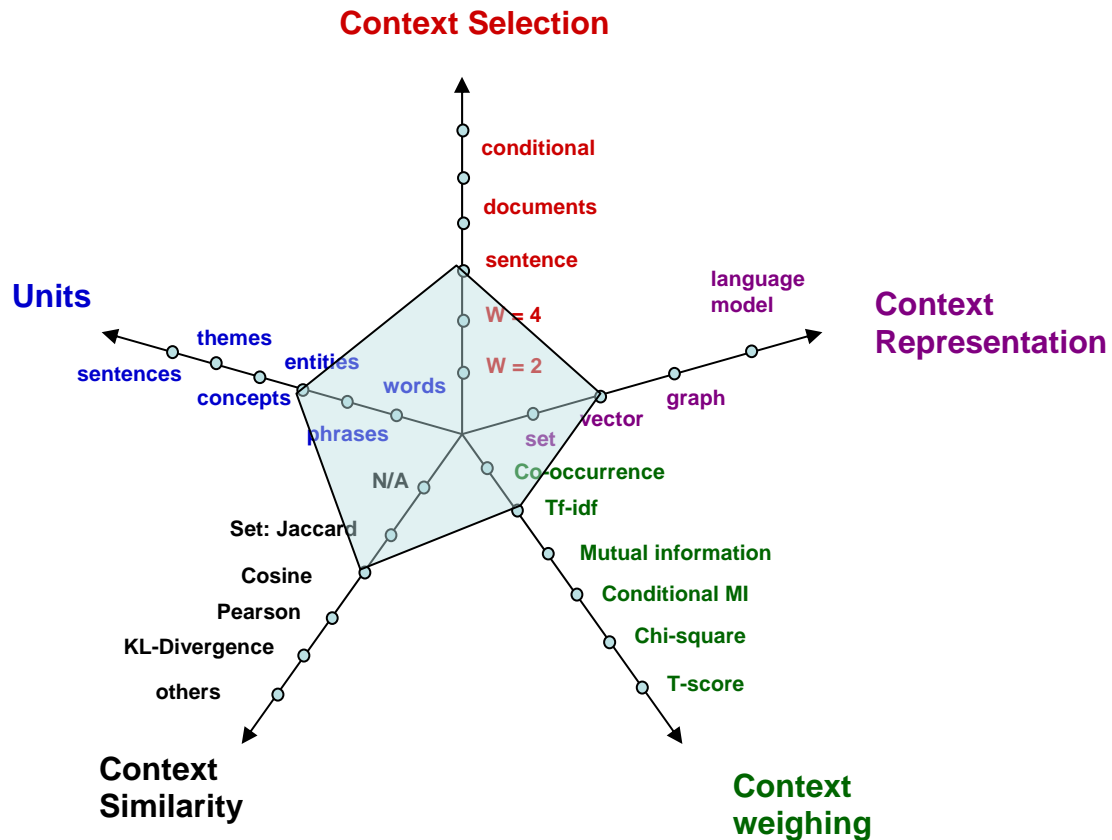
- Problem:
 - ranking
- Source Units
 - term
- Target Units:
 - term
- Context:
 - document
 - Vector
- Context unit weight:
 - Tf-idf, MI
- Context Similarity:
 - N/A

Example: N-gram Clustering



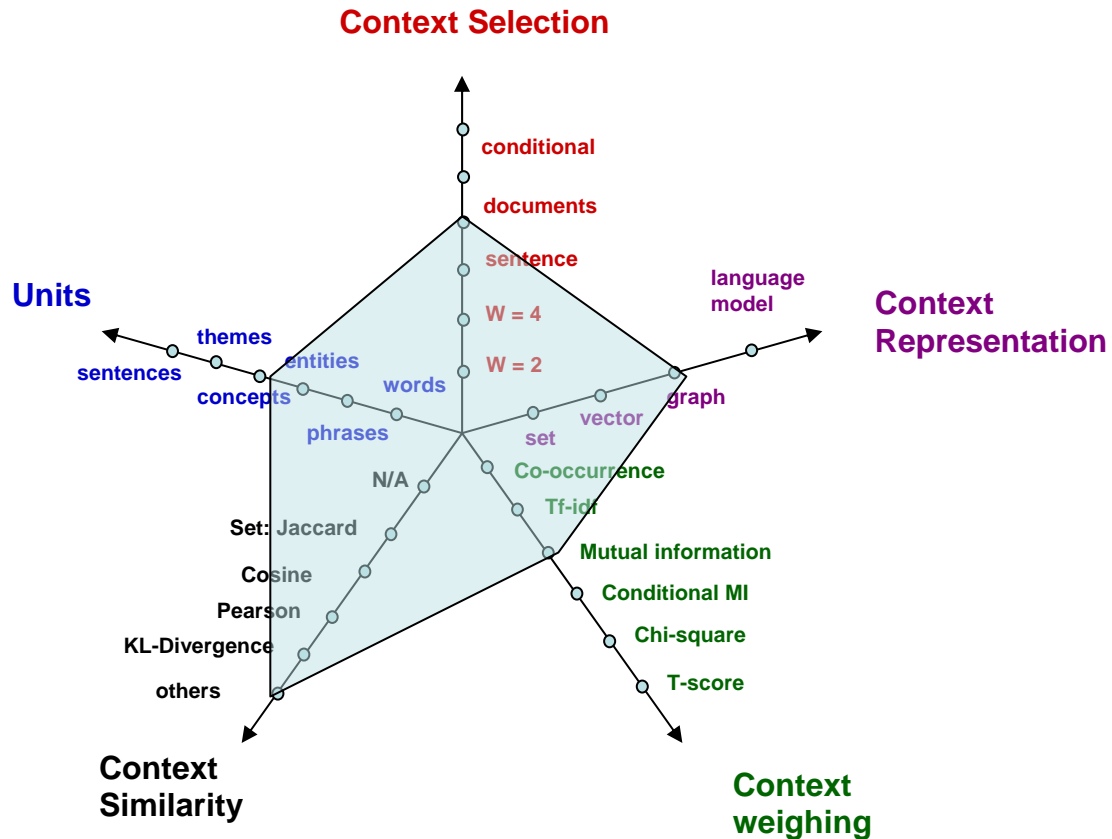
- Problem:
 - clustering
- Source Units
 - Term
- Target Units:
 - group of terms
- Context:
 - Local, $w = 3$
 - Vector
- Context unit weight: MI
- Context Similarity:
 - Drop of total MI if merged

Example: Synonym Extraction (Mei et al. 06)



- Problem:
 - ranking
- Source Units
 - entity
- Target Units:
 - entity
- Context:
 - Natural, sentences
 - Vector
- Context unit weight:
 - Co-occurrence, MI,
- Context Similarity:
 - Cosine

Example: Concept Generation (Brad's)

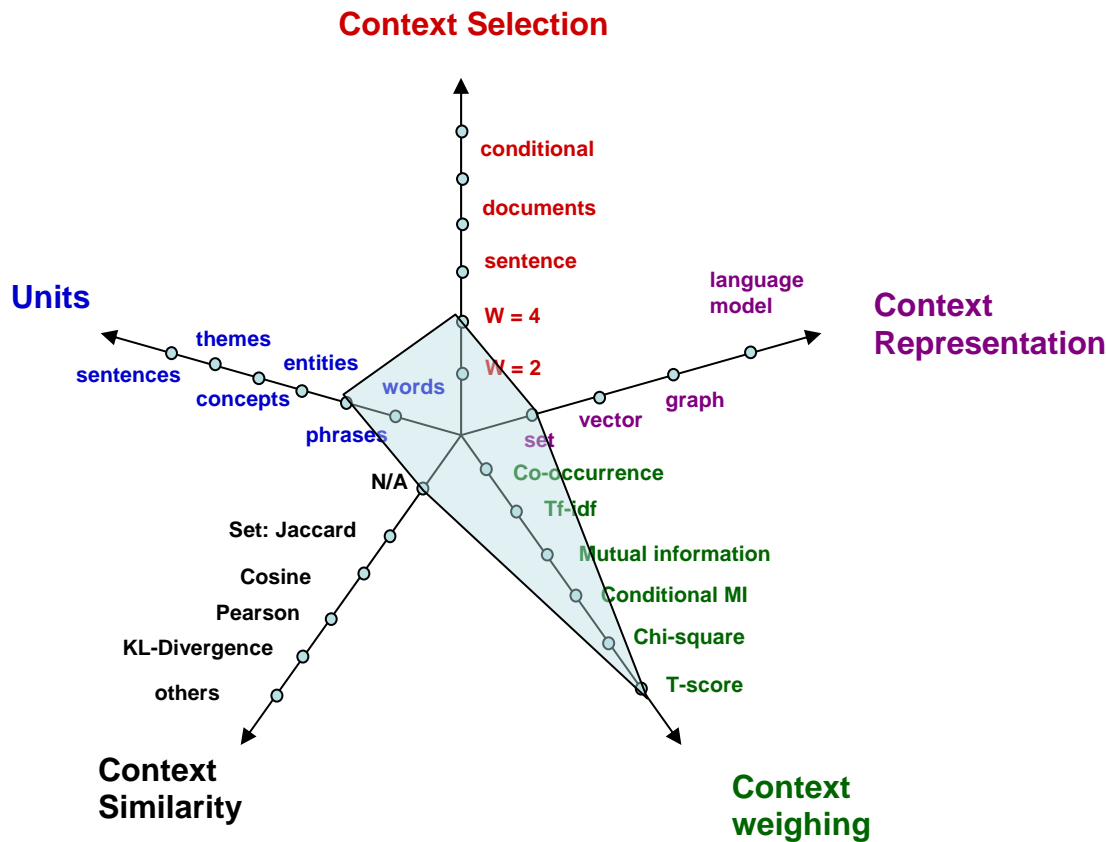


- Problem:
 - clustering
- Source Units
 - Group of words
- Target Units:
 - Group of words
- Context:
 - Documents
 - Graph
- Context unit weight:
 - MI
- Context Similarity:
 - Utility function

Example: Theme Labeling

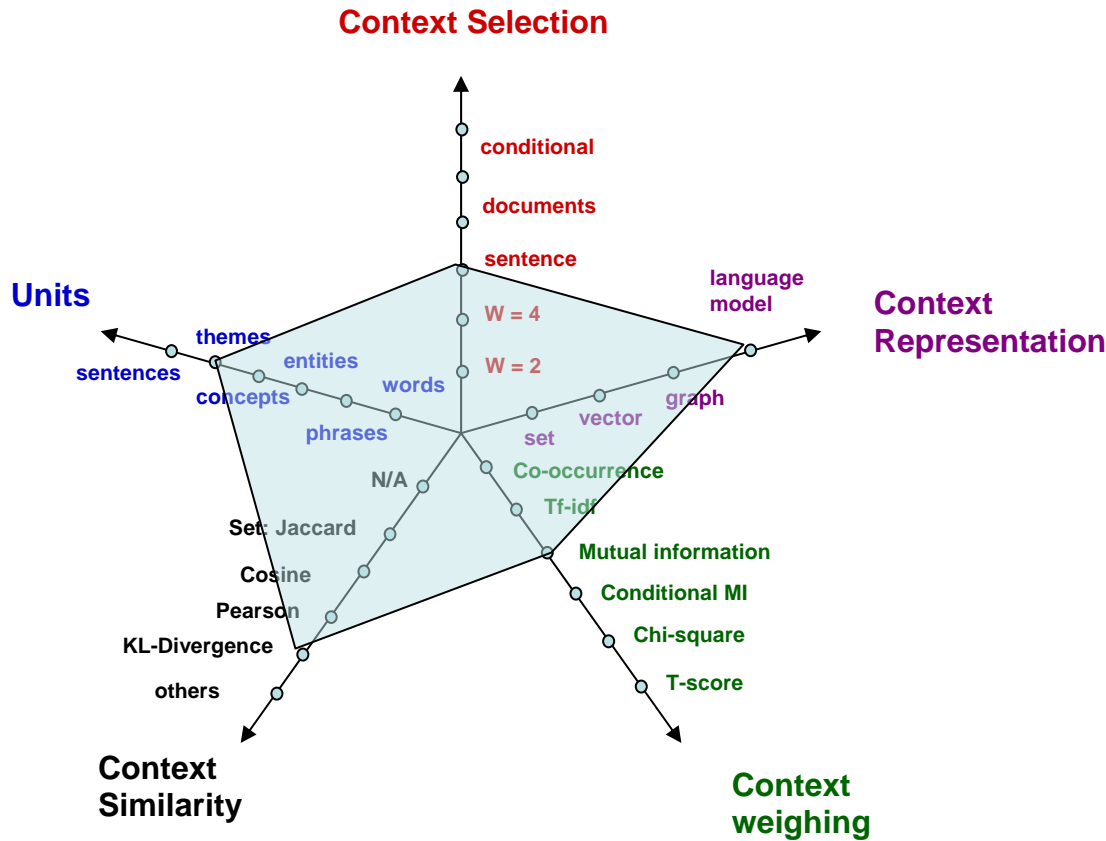
- Source Units: themes
- Target Units: phrases
- Two step:
 - phrase generation → label selection

Theme Labeling: Phrase Generation



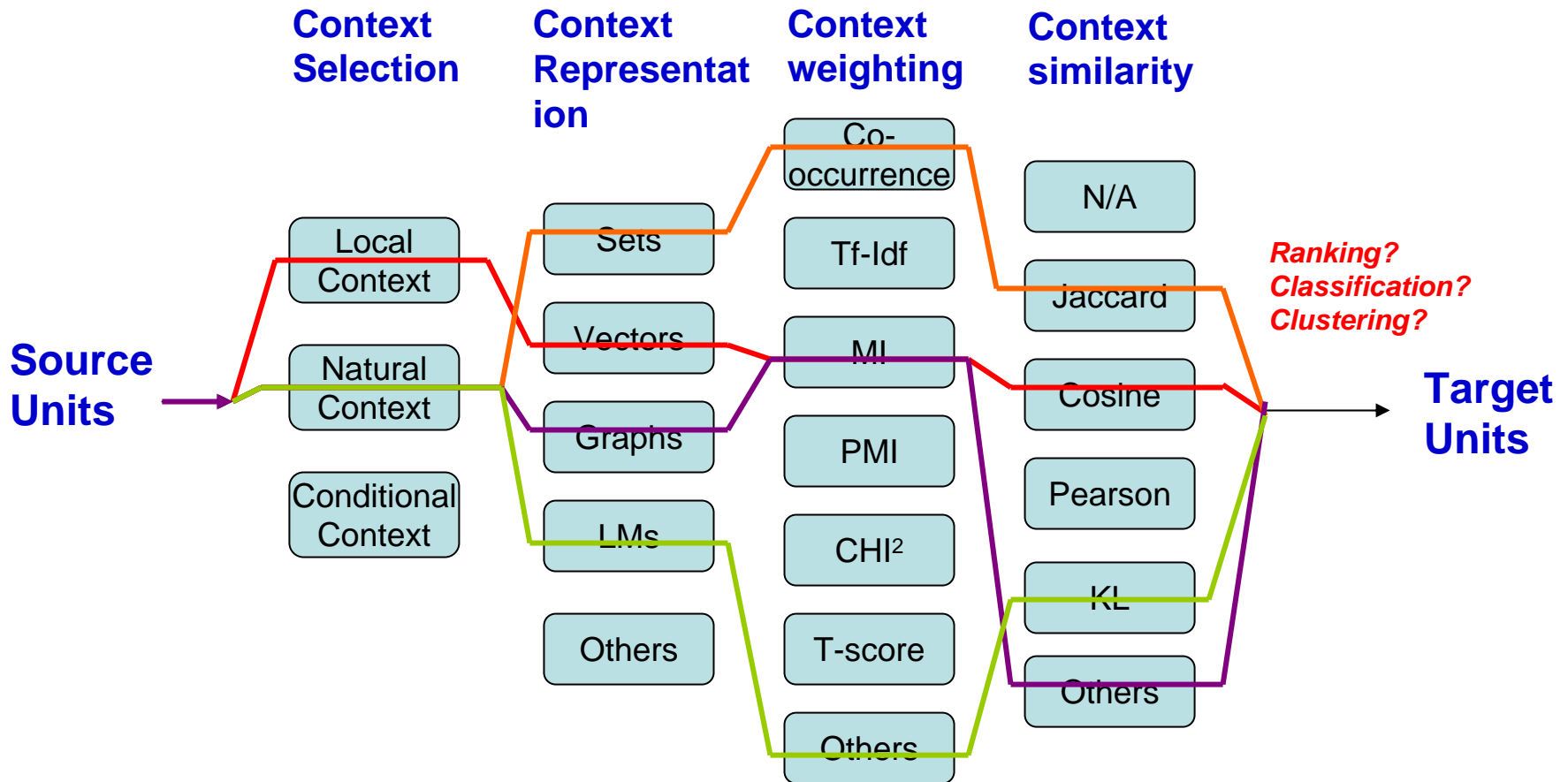
- Problem:
 - ranking
- Source Units
 - words
- Target Units:
 - phrases
- Context:
 - 2-grams
 - pairs
- Context unit weight:
 - MI, t-score, etc
- Context Similarity:
 - N/A

Theme Labeling: Label Selection



- Problem:
 - ranking
- Source Units
 - themes
- Target Units:
 - phrases
- Context:
 - Sentences/docs
 - Language models
- Context unit weight:
 - PMI
- Context Similarity:
 - KL Divergence

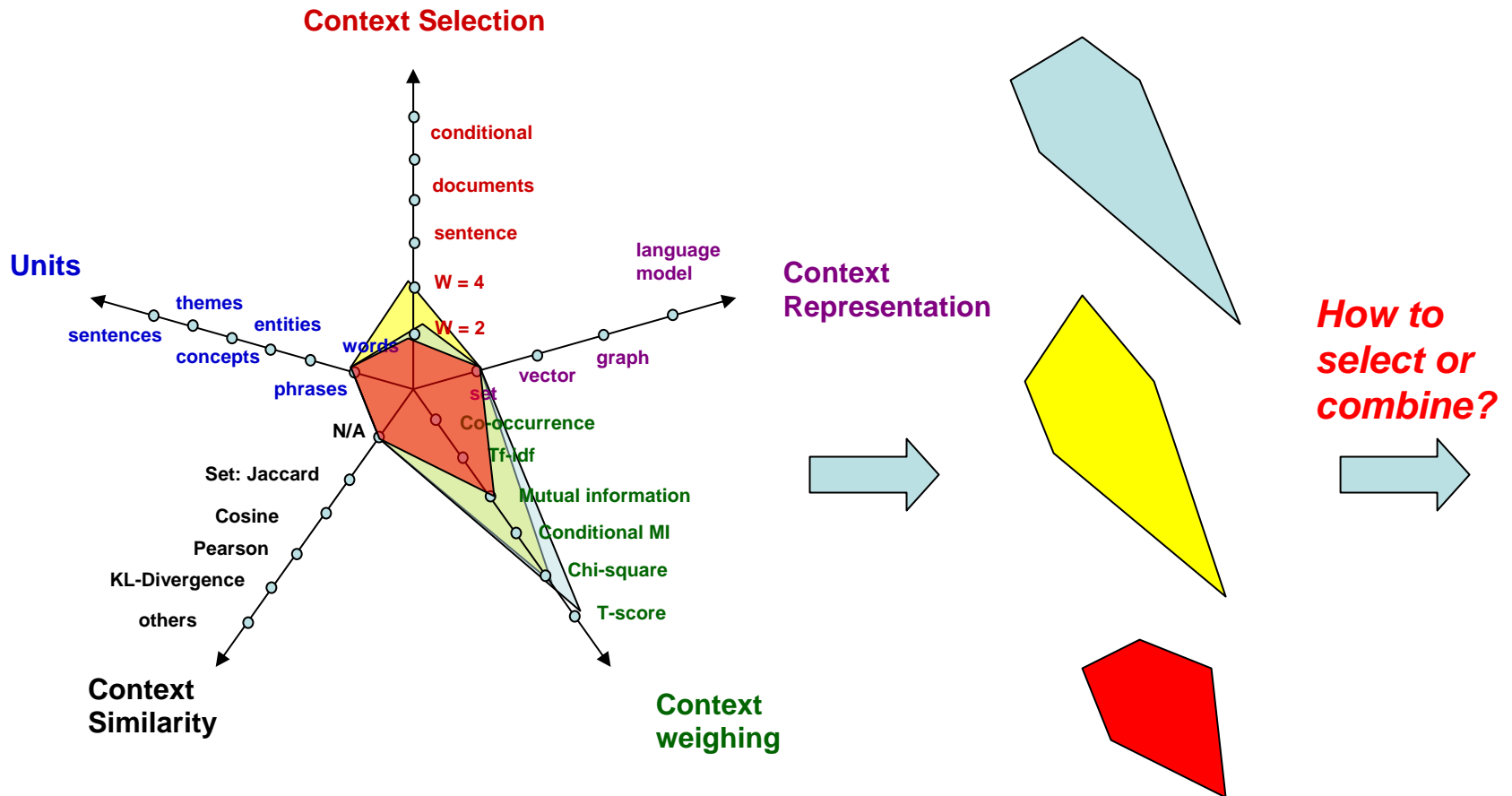
A General Procedure of Semantic Process Problems



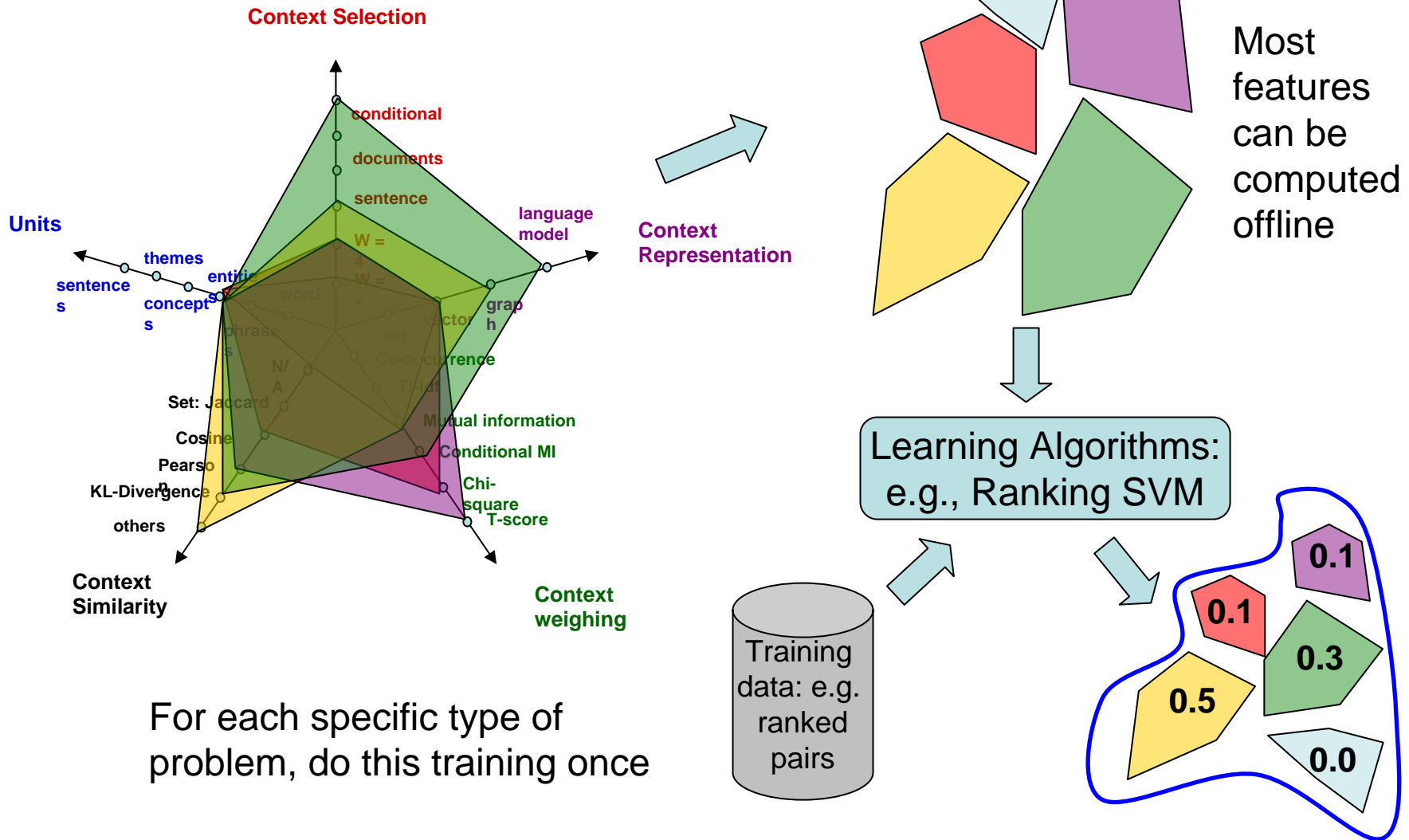
What's Remaining?

- Problem:
 - Too many possible paths from the source units to the target units
 - Given a real problem: each path is a possible solution.
 - Which path in the space is better for a specific problem?
 - Few real problems can be satisfied with one path.
- A possible solution:
 - Model each path as a feature
 - Use training/validation data and learning algorithms to choose good features/combinations of features.

Example: Phrase Generation



Example: Synonym Extraction



Summary

- User select:
 - Type of Source Units; Type of Target Units; Context Conditions
- System Select:
 - Context type; Context representation; Context unit weighting; Context similarity; Weighing/Combination of different features
- Many features can be computed offline;
- For specific selection of user, best combination of features learnt with learning algorithms and training data.