A Modeling Approach for the Extraction of Semantic Information from a Maritime Corpus

Dieudonné Tsatcha, Eric Saux, Christophe Claramunt

Naval Academy Research Institute, GIS group, Lanvéoc-Poullmic, CC 600, F-29240 Brest Cedex 9, France

Symposium on Spatial Data Handling (SDH)
23 August 2012
Plan

- **Introduction**
  - Context
  - Objectives

- **Information Retrieval**
  - Theoretical concepts used in Information Retrieval
  - Capture semantic relation in documents and query
  - Conceptual vector of word

- **Extracting the semantic relation**
  - Conceptual vector of sentence
  - Decision space for the extraction of semantic information
  - Projection of a sentence in a decision space

- **Use case and glosses**
  - Concept anchorage area
  - Cardinal buoy
  - Glosses

- **Conclusion and further work**
Intelligent transport systems (ITS) provide some solutions [Goralski and Gold(2008)]

- Increase the security and the avoidance of hazard area during maritime navigation
- Reduce collisions by improving the visual representation of the environment (ECDIS, Automatic Radar Plotting Aids) [Claramunt et al.(2005)Claramunt, Fournier, Li, and Peytchev]
- Integrate « Intelligent concepts » in the future ECDIS

---

1 Electronic Charts Display and Information System
Objectives

- Extract the knowledge (affordance\(^2\)) generated by an object from a set of maritime documents or corpus.

  - The process is grounded in five steps
    1. Determine concepts derived from maritime navigation documents
    2. Identify the sentences of the corpus related to each concept
    3. Project each sentence of this previous concept in the decision space
    4. Evaluate the contribution of all sentences in the decision space
    5. Compute the direction which has the main contribution, it should be the affordance related to this latter concept.

- Initial corpus contains 16 010 sentences defined with 413 076 words (175 578 nouns, 26 922 verbs, 20 703 adjectives, 9 106 adverbs).

\(^2\)An affordance is a quality of an object, or an environment, which allows an individual to perform an action [Gibson(1977)]
Let us consider the follow sentence $s=\text{“The ship is in the port”}$. The representation of this sentence in the initial corpus is proposed in table 1:

<table>
<thead>
<tr>
<th>word</th>
<th>part-of-speech $^3$</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>DT</td>
<td>the</td>
</tr>
<tr>
<td>ship</td>
<td>NN</td>
<td>ship</td>
</tr>
<tr>
<td>is</td>
<td>VBZ</td>
<td>be</td>
</tr>
<tr>
<td>in</td>
<td>IN</td>
<td>in</td>
</tr>
<tr>
<td>the</td>
<td>DT</td>
<td>the</td>
</tr>
<tr>
<td>port</td>
<td>NN</td>
<td>port</td>
</tr>
</tbody>
</table>

**Table:** Disambiguation of the sentence

$^3$http://en.wikipedia.org/wiki/Brown_Corpus#Partofspeech_tags_used
Objectives

- Propose a safety route to the captain
- Increase captain cognitive abilities during a high stress or high workload situation

Identify entities and taking into account their salience

**Figure:** An example of semantic information that influences the route of vessels by day (on the left) and by night (on the right). By day, the objects are recognized based on their shape and colors whereas they are identified by their light signal by night [Néméta(2008)].
Fondamentally information retrieval (IR) supports three basic processes [Hiemstra (January 2001)]:

- representation of documents content
- representation of query
- comparison of the two previous representations

A suitable model and representation must be determined to extract the semantic relation between documents and query.

The three most used models in IR research are:

1. Vector space model
2. The probabilistic model
3. The inference network model [Singhal (2001)]
Most systems assign a numeric score to every document and rank it.

- These approaches do not take into account the semantic relatedness between query and sentences which satisfy the query.
- [Tsatsaronis and Panagiotopoulou (2009)] points out the importance of capturing semantics between terms in IR.

Proposal algorithm is developed from:

**Algorithm**

- The concept of conceptual vector
- Disambiguation of documents and query. The relevance of results depends on this semantic relatedness.
A conceptual vector of word

- Initially proposed by [Lafourcade et al.(2002)Lafourcade, Prince, and Schwab].

- A conceptual vector of a word is based on the concept of synset of a word.

- Two fundamental elements describe a conceptual vector of word:
  - A set of synsets
  - A metric measuring the distance between this word and each candidate word of a synset

- The conceptual vector of a word is organized according to grammatical categories (adjective (a), adverb (r), noun (n), verb (v)) that the word may belong

- It is also organized according to decreasing distance values inside a grammatical category

---

4 A synset of a word represents a concept and contains a set of interchangeable words, each of them having the same sense that names the concept [Beckwith et al.(1990)Beckwith, Fellbaum, Gross, and Miller]
A conceptual vector of word

Mathematical definition of conceptual vector $V(w)$ of word $w$

$$V(w) = \left( \bigcup_{l=1}^{m^a} c_i^a \delta(c_i^a, w) \right)_a \left( \bigcup_{l=1}^{m^r} c_i^r \delta(c_i^r, w) \right)_r \left( \bigcup_{l=1}^{m^n} c_i^n \delta(c_i^n, w) \right)_n \left( \bigcup_{l=1}^{m^v} c_i^v \delta(c_i^v, w) \right)_v$$ (1)

where $S^c = (s_i^c)_{i=1}^{n^c}$ the sets of $n^c$ synsets of $w$ and $C^c = (c_i^c)_{i=1}^{m^c}$ the sets of $m^c$ candidate words of all synsets in category $c$

Example of conceptual vector of *port*

$$V(port) = (larboard[1.00])_a (embrasure[1.00]porthole[1.00]larboard[1.00]interface[1.00])_n$$
The distance $\delta$ between two words $w_1, w_2$ is computed from the WordNet lexical network as follows:

$$
\delta(w_1, w_2) = \begin{cases} 
1 & \text{if } w_1 \text{ and } w_2 \text{ don’t have a common parent} \\
\| \min(d(w_1, Pg), d(w_2, Pg)) \|_{g \in G} + d(Pg, R) & \text{otherwise}
\end{cases}
$$

(Figure): Illustration of the structure of the lexical network used by WordNet.
A conceptual vector of word

The distance $\delta$ between two words $w_1,w_2$ is computed from the WordNet lexical network as follows:

$$\delta(w_1, w_2) = \begin{cases} 
1 & \text{if } w_1 \text{ and } w_2 \text{ don’t have a common parent} \\
\| \min \left( \frac{d(w_1, Pg), d(w_2, Pg)}{\min(d(w_1, Pg), d(w_2, Pg)) + d(Pg, R)} \right) \|_{g \in G} & \text{otherwise}
\end{cases}$$

(2)

Figure: Illustration of the structure of the lexical network used by WordNet.

$\delta(\text{word1, word2})$ between the two words is equal to $\| \frac{1}{1+1} \| = 0.5$. We assume that the two words have the same grammatical category.
Assuming the principle that a sentence is a collection of polysemic words, a conceptual vector of sentence is the direct sum of conceptual vectors of the words of that sentence.

The resulting of each conceptual vector must be normalised.

Let \( s \) be a sentence composed of \( n \) words \( w_i \), then the conceptual vector of the sentence \( s \) is defined by:

\[
V(s) = \sum_{i=1}^{n-1} V(w_i) \oplus V(w_{i+1})
\]  

(3)

**Definition of direct sum of two words**

\[
V(w_i) \oplus V(w_j) = \begin{cases}
    \bigcup_{k=i,j} \left( \bigcup_{l=1}^{m^a_{w_k}} c^a_{l,w_k} \delta(c^a_{l,w_k}, w_k) \right) & a \\
    \bigcup_{k=i,j} \left( \bigcup_{l=1}^{m^n_{w_k}} c^n_{l,w_k} \delta(c^n_{l,w_k}, w_k) \right) & n \\
    \bigcup_{k=i,j} \left( \bigcup_{l=1}^{m^r_{w_k}} c^r_{l,w_k} \delta(c^r_{l,w_k}, w_k) \right) & r
\end{cases}
\]  

(4)
Conceptual vector of sentence

\[ V(\text{ship}) = (\text{transport}[0.00] \]
\[ \quad \text{send}[0.00] \text{embark}[0.00])_v \]
\[ V(\text{port}) = (\text{larboard}[1.00])_a \]
\[ (\text{embrasure}[1.00] \text{porthole}[1.00] \text{larboard}[1.00] \text{interface}[1.00])_n \]

Direct Sum

\[ V(\text{ship}) \oplus V(\text{port}) = (\text{embrasure}[1.00] \]
\[ \quad \text{porthole}[1.00] \text{larboard}[1.00] \text{interface}[1.00])_n \]
\[ (\text{larboard}[1.00])_a (\text{transport}[0.00] \]
\[ \quad \text{send}[0.00] \text{embark}[0.00])_v \]
We introduce the concept of decision space where the different conceptual vectors of the sentences that describe a given concept must be projected.

The possible options (axes of decision space) should be linked to the objectives and validated by experts.
Projection of a sentence in a decision space

- Allow to valuate the contribution of a sentence in each direction of the decision space
- The projection of a sentence \( s \) in a decision space corresponds to the projection of the candidate words of the conceptual vector of \( s \) (i.e., \( V(s) \)).
- The contribution \( x^c_i \) of a candidate word of \( V(s) \) in the semantic direction \( d \) and category \( c \) is computed as follows:

\[
x^c_i = \frac{(1 - \delta_s \ast \delta_d)}{n}.
\]

where:

- \( n \) is the number of candidate words of the conceptual vector
- \( c_i^c \) is the common candidate word of the two conceptual vectors \( V(s) \) and \( V(d) \) with weights \( \delta_s \) and \( \delta_d \) respectively
The contribution of a sentence in a category $c$ is the sum of the contribution of the $m_c$ candidate words of this category, i.e.,

$$x^c = \sum_{i=1}^{m_c} x^c_i$$

Finally, the contribution of a sentence is the sum of the contributions in each category, i.e., $x = x^a + x^r + x^n + x^v$.

The higher the semantic contribution in a direction, the higher the projection value $x$. 
Projection of a sentence in a decision space

Conceptual vector of word stop
\( V(\text{stop}) \)

(halt[0.00]block[0.00]check[0.00]
arrest[0.00]blockade[0.12] bar[0.14]
end[0.14] finish[0.14]barricade[0.22]
break[0.29]cease[0.67]intercept[0.73]kibosh[1.00]
terminate[1.00]contain[1.00]quit[1.00]discontinue[1.00]) \( V \)

(halt[0.00]stoppage[0.00]stopover[0.00]
layover[0.00]arrest[0.00]check[0.00]hitch[0.00]stay[0.00]
occlusive[0.00]plosive[0.00]period[0.00]point[0.00]
diaphragm[0.00]catch[0.00]blockage[0.00]block[0.00]
closure[0.00]occlusion[0.00]) \( n \)

Conceptual vector of word stay
\( V(\text{stay}) \)

(remain[0.00]rest[0.00]stick[0.00]
bide[0.00]abide[0.00]continue[0.00]detain[0.00]delay[0.00]
persist[0.00]outride[0.00]quell[0.00]appease[0.00]) \( V \)

(arrest[1.00]check[0.33]halt[0.33]stop[0.33]hitch[0.50]
stoppage[1.00]) \( n \)

\( c_1^n=\text{arrest}, c_2^n=\text{check}, c_3^n=\text{halt}, c_4^n=\text{hitch}, 
\)
\( c_5^n=\text{stoppage} \) are the common candidate word between
\( V(\text{stop}) \) and \( V(\text{stay}) \)

\[ x_1^n = \frac{1 - 0.00 \times 1}{35} = 0.029, \]

\[ x^n = x_1^n + x_2^n + x_3^n + x_4^n + x_5^n = 0.15 \]

\[ x = x^a + x^r + x^n + x^v = 0.15 \] and
\( x^a = 0, x^r = 0, x^v = 0 \)
Principal semantic direction for a concept

- Allow to determine the principal semantics in a decision space that is associated to a concept.

**Principal semantic direction is done in three steps:**

- Identify the sentences of the corpus related to this concept and project each of them in the decision space
Principal semantic direction for a concept

- Allow to determine the principal semantics in a decision space that is associated to a concept.

**Principal semantic direction is done in three steps:**

- Identify the sentences of the corpus related to this concept and project each of them in the decision space
- Compute the contribution of each sentence in the decision space
Principal semantic direction for a concept

- Allow to determine the principal semantics in a decision space that is associated to a concept.

**Principal semantic direction is done in three steps:**

- Identify the sentences of the corpus related to this concept and project each of them in the decision space.
- Compute the contribution of each sentence in the decision space.
- Determine the semantic direction which warrants the highest trust (i.e. the one which has the highest coordinate or score).
Allow to determine the principal semantics in a decision space that is associated to a concept.

**Principal semantic direction is done in three steps:**

- Identify the sentences of the corpus related to this concept and project each of them in the decision space
- Compute the contribution of each sentence in the decision space
- Determine the semantic direction which warrants the highest trust (i.e. the one which has the highest coordinate or score)

In the case where at least two directions have the same score, the direction which ensures security of mariner is considered.
Concept “anchorage area”

- Five sentences related to this concept in our corpus.
- We restrict our case study to the actions or affordances (continue, stop, go back and maneuver).
- After projection, the principal affordance we can associate to the concept “anchorage area” is stop with value $0.06 = 0.01 + 0.00 + 0.00 + 0.01 + 0.04$.

**Figure**: Visualisation of the decision space.
Cardinal buoy

Concept “cardinal buoy”

- Two sentences related to this concept in our corpus
- We restrict our case study to the actions or affordances (continue, stop, go back and maneuver)
- After projection, the principal affordance we can associate to the concept “cardinal buoy” is **maneuver** with value $0.04 = 0.02 + 0.02$

**Figure**: Visualisation of the decision space
The glosses of a word are the different definitions of it

The projection of a word or a sentence in a decision space can generate a conceptual vector whose euclidean norm is equal to zero.

Each semantic contribution is equal to zero and no decision emerges

For example, a gloss of a word “port” may be “a place (seaport or airport) where people and merchandise can enter or leave a country”
### Strategy

- We compute the new coordinate of a sentence whose conceptual vector is null by using the glosses of each word in this sentence.

- For each word $w$, the different glosses are extracted and projected in our decision space.

- We use the infinity norm ($\|\|_\infty$) to find the most contributing gloss of a word.
Conclusion

This work introduces a general strategy to extract semantic information from a corpus based on:

- A set of documents written by the experts;
- WordNet, a lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets).

The semantic richness of the initial corpus is important and influences the success of the strategy.

Two strategies are introduced

- The first uses only the synsets
- The second improves the results using the glosses

This semantic information extracted can improve captain cognitive abilities during a high stress or high workload situation.
Further work concerns the development of a real time navigation aid platform which takes into account semantic information generated by objects which appear in the vicinity of the ship.

This platform will be coupled with a spatio-temporal ontology of the maritime environment that will store the initial and the extracted knowledge.
Thank you for your attention
Introduction to WordNet: An Online Lexical Database.

C. Claramunt, S. Fournier, X. Li, and E. Peytchev.
Real-time Geographical Information for ITS.

J. J. Gibson.
The Theory of Affordances, pages 67–82.

R. Goralski and C. Gold.
Marine GIS: Progress in 3D Visualization for Dynamic GIS.

D. Hiemstra.
Using Language Models for Information Retrieval.

M. Lafourcade, V. Prince, and D. Schwab.
Vecteurs conceptuels et structuration émergente de terminologie.

A. Néméta.

A. Singhal.
Modern information retrieval : A brief overview.

G. Tsatsaronis and V. Panagiotopoulou.
A generalized vector space model for text retrieval based on semantic relatedness.
The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), April 2009.