Neologisms

Harvesting & Understanding

Marcel Köster

06/08/2010

Introduction

- widly spread and often used in spoken language before listed in a dictionary
- internet helps the propagation of new words (neologisms)
- Wikipedia
- language processing is hard

Nelogisms created using Variation

- "bloody Mary"
  - tomato juice
  - vodka
- "virgin Mary"

Nelogisms created using Variation

- "bloody Mary"
  - tomato juice
  - vodka
- "virgin Mary"
  1. no tomato juice
  2. no alkohol

Nelogisms created using Variation

- "bloody Mary"
  - tomato juice
  - vodka
- "virgin Mary"
  1. no tomato juice
  2. no alkohol

Nelogisms created using Variation

- "bloody Mary"
  - tomato juice
  - vodka
- "virgin Mary"
  1. no tomato juice
  2. no alkohol
- "Ghost town"
  - a town which has become deserted
- "Ghost airport"

Nelogisms created using Variation

- "bloody Mary"
  - tomato juice
  - vodka
- "virgin Mary"
  1. no tomato juice
  2. no alkohol
- "Ghost town"
  - a town which has become deserted
- "Ghost airport"
  - an airport which has become deserted

Nelogisms created using Combination

- Tourtal

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?
- Tourtal is a nice extension to the list of available games [...]

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?
- Tourtal is a nice extension to the list of available games [...]
  1. Tourtal is game with a Turtle / Toirtoise
  2. ... ?

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?
- Tourtal is a nice extension to the list of available games [...]
  1. Tourtal is game with a Turtle / Toirtoise
  2. ... ?
- ... for Microsoft Surface.

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?
- Tourtal is a nice extension to the list of available games [...]
  1. Tourtal is game with a Turtle / Toirtoise
  2. ... ?
- ... for Microsoft Surface.
  1. Microsoft Surface is a multitouch-table
  2. Portal developed by Valve

Nelogisms created using Combination

- Tourtal
  1. Toirtoise / Turtle
  2. ... ?
- Tourtal is a nice extension to the list of available games [...]
  1. Tourtal is game with a Turtle / Toirtoise
  2. ... ?
- ... for Microsoft Surface.
  1. Microsoft Surface is a multitouch-table
  2. Portal developed by Valve
- "Touchtable-Portal"
- $\Rightarrow$ Tourtal is a Touchtable-version of the game Portal

Nelogisms created using Variation and Combination

- Combination & Variatation are common "tools" in creative language
- How can we detect and understand neologisms?
  - ... where does the background knowledge come from?
  - ... where do the neologisms come from?
  - ... how can we recognize a neologism?
  - ...

Zeitgeist

Idea

use Wikipedia to extract Neologisms and feed them into WordNet

- rule-based approach (instead of a statistical one)
- restricted to "portmanteau" words
  - "two meanings packed up into one word"

Wikipedia → WordNet

- easy to model semantic relations
- *isa* Relation

  if X *isa* $Y \Rightarrow Y$ is a generalization of X
  watergate *isa* gate (is a gate opening onto water)

- *hedges* Relation

  if X *hedges* $Y \Rightarrow X$ ~~*isa*~~ Y but X shares properties with Y
  "kilobit" ~~*isa*~~ "kilobyte" but shares attributes like:
  - relative size "kilo"
  - related to the binary system

Zeitgeist structure

1. Detect neologisms without any knowledge
2. Detect neologisms using knowledge from Pass 1
3. All neologisms detected and understood

Notations & Definitions

- string-matching approach
- $\alpha\beta$ is a general form of a Wikipedia article ("watergate")
- $\alpha \rightarrow \beta$
  (Hardware $\rightarrow$ Electronics)
- $\alpha \rightarrow \beta \; ; \gamma$
  (Electronics $\rightarrow$ Transmitter, Electronic Circuit)
- $\frac{\textit{condition}}{\textit{conclusion}} \quad \frac{\alpha \rightarrow \beta}{\gamma}$

Zeitgeist Pass 1 - learning from easy cases

Schema 1: Explicit extension

$$\frac{\alpha\beta \rightarrow \beta \wedge \alpha\beta \rightarrow \alpha\gamma}{\alpha\beta \ \textit{isa} \ \beta}$$

1. Input: "gastropub"
2. Split the word: $\alpha =$ "gastro", $\beta =$ "pub"
3. "pub" is a valid article $\Rightarrow \alpha\beta \rightarrow \beta$ is fullfilled

Zeitgeist Pass 1 - learning from easy cases

Schema 1: Explicit extension

$$\frac{\alpha\beta \rightarrow \beta \wedge \alpha\beta \rightarrow \alpha\gamma}{\alpha\beta \ \textit{isa} \ \beta}$$

1. Input: "gastropub"
2. Split the word: $\alpha =$ "gastro", $\beta =$ "pub"
3. "pub" is a valid article $\Rightarrow \alpha\beta \rightarrow \beta$ is fullfilled
4. "gastro" is a prefix of "gastronomy" - $\gamma =$ "nomy"
5. gastropub is a pub

Zeitgeist Pass 1 - learning from easy cases

Schema 2: Suffix alternation

$$\frac{\alpha\beta \rightarrow \alpha\gamma \wedge \beta \rightarrow \gamma}{\alpha\beta \textit{ hedges } \alpha\gamma}$$

1. Input: "gigabyte"
2. Split the word: $\alpha =$ "giga", $\beta =$ "byte"
3. "gigabit", $\alpha =$ "giga", $\gamma =$ "bit"
4. "byte" $\rightarrow$ "bit" ($\beta \rightarrow \gamma$ fullfilled)
5. "gibabyte" has something to do with "gigabit"

Zeitgeist Pass 1 - learning from easy cases

Schema 3: Partial suffix

$$\frac{\alpha\beta \rightarrow \gamma\beta \wedge (\alpha\beta \rightarrow \alpha \vee \alpha\beta \rightarrow \delta \rightarrow \alpha)}{\alpha\beta \textit{ hedges } \gamma\beta}$$

1. Input: "software"
2. Split the word: $\alpha =$ "soft", $\beta =$ "ware"
3. $\gamma =$ "computational-application-" $\beta =$ "ware"
4. "software" has a reference to "computational-application-ware" ($\alpha\beta \rightarrow \gamma\beta$ fullfilled)
5. "software" has a reference to "soft" ($\alpha\beta \rightarrow \alpha$ fullfilled)
6. "software" is related to "computational-application-ware"

Zeitgeist Pass 1 - learning from easy cases

Schema 4: Consecutive Blends

$$\frac{\alpha\beta \rightarrow \alpha\gamma; \delta\beta}{\alpha\beta \ \textit{hedges} \ \delta\beta}$$

1. Input: "sharpedo"
2. Split the word: $\alpha =$ "shar", $\beta =$ "pedo"
3. $\gamma =$ "k" $\rightarrow \alpha\gamma =$ "shark"
4. $\delta =$ "tor" $\rightarrow \delta\beta =$ "torpedo"
5. "sharpedo" has reference to "shark" and "torpedo"
6. "sharpedo" is related to a "torpedo"

Zeitgeist Pass 1 - learning from easy cases

Schema 4 $\frac{1}{2}$: The obvious case

$$\frac{\alpha\beta \rightarrow \gamma \ ; \ \delta \ (\textit{portmanteau})}{\alpha\beta \ \textit{hedges} \ \gamma \wedge \alpha\beta \ \textit{hedges} \ \delta}$$

1. Input: "spork"
2. Zeitgeist recognizes extension "portmanteau-word"
3. Extract $\gamma =$ "spoon", $\delta =$ "fork"
4. "spork" is related to "spoon" and "fork"

Zeitgeist Pass 1 - summary

Schema	Word
Explicit extension	"gastropub"
Suffix alternation	"gigabyte"
Partial suffix	"software"
Consecutive Blends	"sharpedo"
The obvious case	"spork"

Zeitgeist Pass 2 - resolving opaque cases

Schema 5: Suffix Completion

$$\frac{\alpha\beta \rightarrow \gamma\beta \wedge \gamma\beta \in E \wedge \beta \in S}{\alpha\beta \textit{ hedges } \gamma\beta}$$

$E :=$ set of all analysed words from rules 3 and 4 (software)

$S :=$ corrseponding set of partial suffixes (ware)

1. Input: "middleware", $\alpha =$ "middle", $\beta =$ "ware"
2. has a reference to "software" ($\alpha\beta \rightarrow \gamma\beta$ fullfilled)
3. "software" is known from schema 3 ($\beta \in E$ fullfilled)
4. "ware" is a valid partial suffix($\beta \in S$ fullfilled)
5. "middleware" is related to "software"

Zeitgeist Pass 2 - resolving opaque cases

Schema 6: Seperable Suffix

$$\frac{\alpha\beta \rightarrow \beta \wedge \alpha \in P}{\alpha\beta \ \textit{isa} \ \beta}$$

$P :=$ set of all prefixes identified by rules 1, 2 and 3 (giga-, soft-)

1. Input: "antiprism"
2. Split the word: $\alpha =$ "anti", $\beta =$ "prism"
3. "antiprism" has a reference to "prism" ($\alpha\beta \rightarrow \beta$ is fullfilled)
4. "anti" is known from schema 1 ($\alpha \in P$ is fullfilled)
5. "antiprism" is a "prism"

Zeitgeist Pass 2 - resolving opaque cases

Schema 7: Prefix Completion

$$\frac{\alpha\gamma \rightarrow \alpha \wedge <\gamma, \delta\beta> \in T}{\alpha\beta \ \textit{isa} \ \beta}$$

$T :=$ set of all tuples identified by rule 1 (<gastro, pub>)

1. Input: "restaurantgastro"
2. Split the word: $\alpha =$ "restaurant", $\gamma =$ "gastro"
3. "restaurantgastro" has a reference to "restaurant" ($\alpha\gamma \rightarrow \alpha$ fullfilled)

Zeitgeist Pass 2 - resolving opaque cases

Schema 7: Prefix Completion

$$\frac{\alpha\gamma \rightarrow \alpha \wedge <\gamma, \delta\beta> \in T}{\alpha\beta \ \textit{isa} \ \beta}$$

$T :=$ set of all tuples identified by rule 1 (<gastro, pub>)

1. Input: "restaurantgastro"
2. Split the word: $\alpha =$ "restaurant", $\gamma =$ "gastro"
3. "restaurantgastro" has a reference to "restaurant" ($\alpha\gamma \rightarrow \alpha$ fullfilled)
4. <gastro, pub> $\in T$, $\delta = \emptyset$, $\beta =$"pub"
5. "restaurantpub" isa "pub"

Zeitgeist Pass 2 - resolving opaque cases

Schema 8: Recombination

$$\frac{\alpha\beta \rightarrow \alpha\gamma \wedge \alpha\beta \rightarrow \delta\beta \wedge \alpha \in P \wedge \beta \in S}{\alpha\beta \textit{ hedges } \delta\beta}$$

1. Input: "geonym"
2. Split the word: $\alpha =$ "geo", $\beta =$ "nym"
3. "geo" is valid prefix from pass 1 ($\alpha \in P$ fullfilled)
4. "nym" is valid suffix from pass 1 ($\beta \in S$ fullfilled)
5. "geonym" has a reference to "geography" ($\alpha\beta \rightarrow \alpha\gamma$ fullfilled)
6. "geonym" has a reference to "toponym" ($\alpha\beta \rightarrow \delta\beta$ fullfilled)
7. "geonym" stands in relation to "toponym"

Zeitgeist Rules

Schema	Word
Explicit extension	"gastropub"
Suffix alternation	"gigabyte"
Partial suffix	"software"
Consecutive Blends	"sharpedo"
The obvious case	"spork"
Suffix Completion	"middleware"
Seperable Suffix	"antiprism"
Prefix Completion	"restaurantpub" ("restaurantgastro")
Recombination	"geonym"

Evaluation

- analysed 152.600 potential neologism words
- 4677 are detected using one or more rules
- 2269 ignored
- remaining 51% (2408) were analysed

Schema	#	Words	# Errors	Precision
Schema 1: Explicit extension	710	(29%)	11	0.985
Schema 2: Suffix alternation	144	(5%)	0	1.0
Schema 3: Partial suffix	330	(13%)	5	0.985
Schema 4: Consecutive Blends	82	(3%)	2	0.975
Schema 5: Suffix Completion	161	(6%)	0	1.0
Schema 6: Seperable Suffix	321	(13%)	16	0.95
Schema 7: Prefix Completion	340	(14%)	32	0.9
Schema 8: Recombination	320	(13%)	11	0.965

Conclusion

1. Pro
   - usage of Wikipedia as
     - background-knownledge database
     - source "corpus"
   - usage of WordNet to model semantic dependencies
   - rule-based approach to match portmanteau-words
   - ... ?
2. Contra
   - disambiguation features missing
   - Wikipedia-dependent
   - ... ?

Thank You

Thanks for your attention :-)

Questions?

References

1. Veale, Butnariu (2010). Harvesting and understanding on-line neologisms
2. Deleuze, Gilles (1990). The logic of sense
3. Miller, George (1995). WordNet: A Lexical Database for English
4. Ruiz-Casado et. al (2005b). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet