

Portmanteau Blends Neologisms WordNet

Cultural Learnings of Wikipedia
For make benefit glorious resource of WordNet

Zeitgeist

Words Wikipedia

Tony Veale
School of Computer Science,
UCD (NUI-D)
Tony.Veale@UCD.ie

Gestalts
Affixes
Prefixes

What is the Secret
of ...
BRIDEZILLA ?!?



Neologisms and the Lexicon

- Neologisms are novel word forms (New-Word)

Definition

E.g., "soccer mom", "neocon", "gastropub", "influenza", "chicken hawk"

- Neologisms enter the lexicon gradually

Bubbling Under

Can be used without gaining widespread recognition for years

- Wikipedia

Reflects Cultural Change

Updated more frequently than print dictionaries, captures the "*Zeitgeist*"

- ZeitGeist

Neologism Harvester

Trawls Wikipedia looking for neologisms (w.r.t WordNet) and their meanings

Lexicographic Approaches to Neologism Analysis

- **Explanatory Lexicography**

Post-Hoc Analysis

Seeks to dissect and explain neologisms after they have been identified in text

- **Predictive Lexicography**

Pre-Hoc Creativity

Seeks to predict and verify neologisms using principles of word-formation

- **Explanatory Lexicography at Work**

Example #1

"Affluenza" = "Affluent" + "Influenza" (overlap: "...*fluen*...")

- **Predictive Lexicography at Work**

Example # 2

"Chrono-" (= time) + "-onaut" (=traveller) = "chrononaut" (a "time traveller")

Trawling Wikipedia

- **Uni-Directional Links**

Notation: $A \longrightarrow B$

Headwords accompanied by text article that links to relevant other articles

- **Bidirectional Reciprocated Links**

Notation: $A \longleftrightarrow B$

Two headwords cross-reference each other, indicating strong mutual association

- **Contiguous Links**

Notation: $A \longrightarrow B ; C$

The Wikipedia article for *A* contains a link to *B* followed directly by a link to *C*

- **Text-Free Analysis (No Parsing Required)**

Topological Approach

The text of each article is ignored; only topology of headword xrefs is used

Compound Terms

- **Simplest: ADJ-Noun and Noun-Noun collocations** **Prevalent in English/ PWN**

Yokes two branches of an ontology together: e.g., Religious-Music, Chinese-Cuisine

- **General Structure**

Endocentricity

<Modifier : Head>, where Head denotes a hypernym of the compound meaning

- **Awkward Exceptions**

Exocentricity

<Modifier : Head> does not denote a hyponym of {Head}, e.g., "hammer head"

- **Inter-Compound Relationships**

Specialization of Parts

Applied-Science :: Applied-Physics but Religious-Music :: Christian-Music

Compound Schema I : Head Specialization

Endocentric form: α β

$$\frac{\alpha_{\beta} \rightarrow \alpha_{\gamma} \quad \wedge \quad \beta \text{ isa } \gamma}{\alpha_{\beta} \text{ isa } \alpha_{\gamma}}$$

*PWN senses of β
and γ given by
this relationship*

- **Examples**

Fantasy_Sport \rightarrow "Fantasy_Football and Football isa Sport (in PWN)

Vetinary_Science \rightarrow Vetinary_Medicine and Medicine isa Science (in PWN)

Compound Schema II : Head Anchoring

Anchored in non-compound

$$\frac{\alpha_\beta \rightarrow \gamma \quad \wedge \quad \beta \text{ isa } \gamma}{\alpha_\beta \text{ isa } \gamma}$$

*PWN senses of β
and γ given by
this relationship*

- **Examples**

Applied_Statistics \rightarrow Science and Statistics isa Science (in PWN)

British_Rail \rightarrow Railway and Rail isa (shorthand for) Railway (in PWN)

Compound Schema III : Compound Expansion

Synonym Creation

$$\frac{\alpha_\beta \rightarrow \alpha \quad \wedge \quad \alpha \text{ isa } \beta}{\alpha_\beta \text{ synonym of } \alpha}$$

*PWN senses of β
and γ given by
this relationship*

- **Examples**

Oscar_award \rightarrow Oscar and Oscar isa award (in PWN)

Semitic_language \rightarrow Semitic and Semitic isa Language (in PWN)

Compound Schema IV : Compound Conflation

Schema linking

$$\alpha_ \beta \rightarrow \beta_ \gamma \quad \wedge \quad \beta \text{ isa } \gamma$$
$$\alpha_ \beta \text{ isa } \beta_ \gamma$$

*$\beta_ \gamma$ may be a
new compound,
previously
learned*

- **Examples**

Touch_rugby \rightarrow Rugby_football and Rugby isa football (in PWN)

Pop_punk \rightarrow Punk_Rock and Punk isa Rock (in PWN)

Compound Schema V : Compound Opposition

Antonym Creation

$$\frac{\alpha_\beta \rightarrow \alpha_\gamma \quad \wedge \quad \gamma \text{ antonym } \beta}{\alpha_\beta \text{ antonym of } \alpha_\gamma}$$

*PWN senses of β
and γ given by
this relationship*

• Examples

Second_language \rightarrow First_language and **Second** antonym of **First** (in PWN)

Synthetic_geometry \rightarrow Analytic_geometry and **Synthetic** opposes **Analytic**

Compound Schema VI : Head Expansion

See also schema III

$\alpha_\beta \rightarrow \gamma_\beta \wedge \gamma_\beta \text{ synonym of } \beta$

$\alpha_\beta \text{ isa } \gamma_\beta$

*PWN senses of β
and γ_β given
by this*

- **Examples**

Escort_Carrier \rightarrow Aircraft_carrier and Aircraft_carrier syn. of Carrier

Simple_majority \rightarrow Absolute_Majority and Absolute_Majority syn. Majority

Compound Schema VII : Modifier Specialization

+ *vice versa*

$$\frac{\alpha_\beta \rightarrow \gamma_\beta \quad \wedge \quad \alpha \text{ mod-isa } \gamma}{\alpha_\beta \text{ isa } \gamma_\beta}$$

*PWN senses of α
and γ given by
this relationship*

- **Examples**

Truck_racing \rightarrow Auto_racing and **Truck** isa automotive_Vehicle (in PWN)

Hindu_music \rightarrow Religious_Music and **Hindu** isa **Religious_person** in (PWN)

Compound Schema VIII : Catch-all Coordination

Weakest schema

$$\alpha_ \beta \rightarrow \gamma_ \beta \quad \wedge \quad \gamma_ \beta \text{ isa } \beta$$

$\alpha_ \beta$ *coordinate of* $\gamma_ \beta$

*PWN senses of β
and $\gamma_ \beta$ given
by this*

- **Examples**

Financial_mathematics \rightarrow applied_mathematics (isa mathematics)

Constructed_language \rightarrow natural_language (isa language)

Endocentric Compound Evaluation: Results

For 10,899 Wikipedia headwords matching one or more compound schema:

E.g.,

Dutch_language
isa
German_language

E.g.,

Supreme_Court
isa
state_court

Schema	# Headwords	# Error	Precision
I	14%	0	1.0
II	12%	0	1.0
III	13%	0	1.0
IV	4%	0	1.0
V	1%	0	1.0
VI	3%	0	1.0
VII	15%	7%	.93
VIII	70%	0	1.0

In-depth Analysis: Portmanteau Words

- Portmanteau (double-pocket) words

Lewis Carroll

A Textual blend of two different words, e.g., "Bollywood", "Infomercial"

- General Structure

Prefix + Suffix

One words contributes a prefix, the other a suffix (but, e.g., "Modem")

- True Portmanteaux

Double-Scope Blends

Neither component word is present in its entirety, (e.g., "metrosexual")

- Impure Portmanteau

Broad Exceptions

E.g., "Wikipedia", (but not "Wiktionary"), "Gastropub", "Feminazi", ...

Taxonomic Connectives

- **Precise Taxonomic Placement**

Pure ISA

E.g., "Superhero" = "Super-" + "Hero" \Rightarrow *Superhero* ISA *Hero*

- **Approximate Taxonomic Placement**

Hedging

E.g., "Spintronics" = "Spin" + "Electronics" \Rightarrow *Spintronics* hedges *Electronics*

- **Disambiguation**

Sense Priming

E.g., Which sense of "hero" does "Superhero" extend? (not a sandwich obviously)

- **Hedging Supports Figurative Portmanteaux**

Metaphor

"Affluenza" = "Affluence" + "Influenza" but *Affluenza* NOT-ISA *Influenza*

General Approach: Two-Pass Harvesting

A Textual, String-Matching Approach

String-Matching

Let $\alpha\beta$ represent the general form of a headword; analyse with schemata

The Topological Context

Wikipedia

The set of Wikipedia cross-references for a given headword $\alpha\beta$

Pass I: Learning from Easy Cases

Easy

Harvest obvious examples (in rich contexts) first, and learn from these cases

Pass II: Applying learnt patterns to Hard Cases

Hard

When topological context is insufficient, use experience of easy cases as a guide

Portmanteau Schema I : Explicit Extension

Easy

$$\frac{\alpha\beta \rightarrow \beta \quad \wedge \quad \alpha\beta \rightarrow \alpha\gamma}{\alpha\beta \text{ isa } \beta}$$

- **Examples**

Gastropub ("Gastropub" \rightarrow "pub" and "Gastropub" \rightarrow "Gastronomy")

Feminazi ("Feminazi" \rightarrow "Nazi" and "Feminazi" \rightarrow "Feminism")

Portmanteau Schema II : Suffix Alternation

Easy

$$\frac{\alpha\beta \rightarrow \alpha\gamma \quad \wedge \quad \beta \leftrightarrow \gamma}{\alpha\beta \text{ hedges } \alpha\gamma}$$

- Examples

"man" \leftrightarrow "boy" "woman" \leftrightarrow "girl" "bit" \leftrightarrow "byte" "toxin" \leftrightarrow "bacteria"

Fangirl ("Fangirl" \rightarrow "Fanboy" and "boy" \rightarrow "girl")

Portmanteau Schema III : Partial Suffix

Easy

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

- Examples

"Metrosexual" \rightarrow "Heterosexual" \wedge "Metrosexual" \rightarrow "Metro"

"Pomosexual" \rightarrow "Homosexual" \wedge "Pomosexual" \rightarrow "Postmodernism" \rightarrow "pomo"

Portmanteau Schema IV : Consecutive Blending

Easy

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

- Examples

"sharpedo" \rightarrow "shark" ; "torpedo" *hedges* "torpedo"

"Spanglish" \rightarrow "Spanish" ; "English" *hedges* "English"

Portmanteau Schema IVa : Partial Suffix

Easy

$$\alpha\beta \rightarrow \alpha\gamma ; \delta\beta \wedge \alpha\beta \rightarrow \textit{portmanteau}$$

$\alpha\beta$ *hedges* $\gamma\beta$

- **Examples**

"Spork" \rightarrow "Spoon" \wedge "Fork" *hedges* "English" (2 characters)

"Sporgergy" \rightarrow "Spam" \wedge "Forgery" *hedges* "Forgery" (2 characters)

Portmanteau Schema V : Suffix Completion

Hard

$$\frac{\alpha\beta \rightarrow \gamma\beta \wedge \gamma\beta \in \mathcal{H}_{\text{III}} \wedge \beta \in \mathcal{S}_{\text{III}}}{\alpha\beta \text{ hedges } \gamma\beta}$$

$\alpha\beta$ *hedges* $\gamma\beta$

Schema III

$\mathcal{H}_{\text{III}} = \text{headwords}$

$\mathcal{P}_{\text{III}} = \text{prefixes}$

$\mathcal{S}_{\text{III}} = \text{suffixes}$

- **Examples**

"Retrosexual" \rightarrow "Metrosexual" \wedge "sexual" $\in \mathcal{S}_{\text{III}}$

"crippleware" \rightarrow "malware" \wedge "ware" $\in \mathcal{S}_{\text{III}}$

Portmanteau Schema VI : Separable Suffix

Hard

$$\alpha\beta \rightarrow \beta \wedge \alpha \in (\mathcal{P}_I \cup \mathcal{P}_{II} \cup \mathcal{P}_{III})$$

$\alpha\beta$ *isa* β

- **Examples**

"Gastroshop" \rightarrow "shop"

\wedge "gastro" $\in \mathcal{P}_{III}$

"antiprism" \rightarrow "prism"

\wedge "anti" $\in \mathcal{P}_{III}$

Portmanteau Schema VII : Prefix Completion

Hard

$$\alpha\gamma \rightarrow \alpha \wedge \langle \gamma, \beta \rangle \in \mathcal{T}_I$$

Schema I

$$\alpha\beta \text{ isa } \beta$$

$$\mathcal{T}_I =$$

Suffix replacement pairings for schema I

- **Examples**

"Logicism" \rightarrow "logic" \wedge "<ism", "Nazi"> $\in \mathcal{T}_I$ so *Logicnazi* **ISA** *Nazi*

"Psychology" \rightarrow "psycho" \wedge "<ology", "technology"> $\in \mathcal{T}_I$

Portmanteau Schema VIII : Recombination

Hard

$$\frac{\alpha\beta \rightarrow \alpha\gamma \ \wedge \ \alpha\beta \rightarrow \delta\beta \ \wedge \ \alpha \in \mathcal{P}_{\text{III}} \ \wedge \ \beta \in \mathcal{S}_{\text{III}}}{\alpha\beta \text{ hedges } \delta\beta}$$

- Examples

"geonym" \rightarrow "geography" \wedge "geonym" \rightarrow "toponym" (hedges geonym)

"dubtitle" \rightarrow "dubbed" \wedge "dubtitle" \rightarrow "subtitle" (hedges subtitle)

Evaluation: Set-Up

- **# of Atomic Headwords in Wikipedia (June 2005)**

Wikipedia

Wikipedia contains 152,060 atomic headwords at this time of download

- **WordNet Version: 1.6**

WordNet

Few differences were achieved using WordNet 2.1 instead

- **# of Wikipedia Entries Matching a ZeitGeist schema**

Initial Selection

4676 headwords match at least one Zeitgeist schema

- **Metaphor is a useful too for ontological development**

Pre-Filtering

1385 already in WordNet; 1083 analyses yield non-PWN parent or hedge

Portmanteau Evaluation: Results

For 2048 Wikipedia headwords matching one or more Portmanteau schema:

E.g.,

Rubbergate
from
Watergate

Schema	# Headwords		# Error	Precision
I	710	29%	11	.985
II	144	5%	0	1.0
III	330	13%	5	.985
IV	82	3%	2	.975
V	161	6%	0	1.0
VI	321	13%	16	.95
VII	340	14%	32	.90
VIII	320	13%	11	.965

E.g.,

Retrosexual
from
Metrosexual

Conclusions

- A Linguistics-Lite Approach to Neologisms is Feasible

Lightweight

No text parsing or morphological analysis; all relevant morphemes are learned

- Taxonomic Hedging is required

Uncertainty

Word-forms are not deterministic w.r.t. taxonomic placement, approx. needed

- Link Topology offers context-specific insights

Grounding

E.g., Microsurgery → Microscopy + surgery ⇒ "surgery done with a microscope"

- Not biased towards English

Multilingual

Linguistics-lite means no language-bias - applicable to other languages / wikis?