DETECTING INTENTIONAL LEXICAL AMBIGUITY IN ENGLISH PUNS

Mikhalkova E.V., Karyakin Yu.E.
Tyumen State University

Figurative language - a challenge...

Ironic>>

*I just **love** working for 6.5 hours without a break or anything.*

Literal>>

*I literally **love** Stephen A smith haha he's hilarious*

From: The ESWC-17 Challenge on Semantic Sentiment Analysis

SemEval Challenges on Figurative Language

- SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses
- SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter
- **SemEval-2017: Detecting sentiment, humor, and truth:**
  - Sentiment Analysis in Twitter
  - Fine-Grained Sentiment Analysis on Financial Microblogs and News
  - #HashtagWars: Learning a Sense of Humor
  - **Detection and Interpretation of English Puns**
  - RumourEval: Determining rumour veracity and support for rumours

Puns

- a short humorous **genre**, where a word or phrase is used intentionally in two meanings:
  - I used to be a banker, but I lost **interest**.
  - Штирлиц открыл окно. Из окна **дуло**. Штирлиц закрыл окно, и **дуло** исчезло.
- a **means of expression**, the essence of which is to use a word or phrase so that in the given context the word or phrase can be understood in two meanings simultaneously:
  - “Пошли в **тренажирный** зал”
  - Romeo: “Not I, believe me. You have dancing shoes with nimble **soles**; I have a **soul** of lead” (Romeo and Juliet)

The Banker Joke

I used to be a banker, but I lost **interest**.

Curiosity	Profit

Штирлиц открыл окно. Из окна **дуло**. Штирлиц закрыл окно, и **дуло** исчезло.

Веять	Элемент стрелкового оружия

Mining Semantic Fields

Roget's Thesaurus is a widely used English-language thesaurus, created in 1805 by Peter Mark Roget (1779–1869), British physician, natural theologian and lexicographer. It was released to the public on 29 April 1852. The original edition had 15,000 words.

Roget’s Thesaurus is composed of six primary classes. Each class is composed of multiple divisions and then **39 sections**. This may be conceptualized as a tree containing over a thousand branches for individual "meaning clusters" or semantically linked words.

Fields of the Banker Joke

use
24, Volition In General
30, Possessive Relations
be
0, Existence
19, Results Of Reasoning
banker
31, Affections In General
30, Possessive Relations

Stop-words excluded

lose
21, Nature Of Ideas Communicated
26, Results Of Voluntary Action
30, Possessive Relations
19, Results Of Reasoning
interest
30, Possessive Relations
25, Antagonism
24, Volition In General
7, Causation
31, Affections In General
16, Precursory Conditions And Operations
1, Relation

Semantic Vector of the Banker joke

p_{Banker}={1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 1, 0, 0, 2, 1, 1, 0, 0, 0, 4, 2}

Sorted vector, decreasing order:

{4, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}

Split vector:

{1, 1, 1, 0, 0, 0, 0, 0 | 0, 0, 0, 0, 0, 0, 0, 0 | 2, 1, 1, 0, 0, 0, 0, 0 | 4, 2, 2, 1, 1, 0, 0, 0, 0, 0}

Research

Train set: 1,240 puns 1,240 random sentences

Test set: 1,240 puns 1,240 random sentences

	F-score	**Sorted vector**	**Split vector**
Linear SVM	0.66	0.56	0.63
SVM with RBF	0.68	0.58	0.66

Hitting the Target Word

I used to be a banker, but I lost **interest**.

The target word is a word that immediately belongs to two semantic fields.

The target word tends to occur at the end of the sentence.

Word Values

Candidate semantic fields: two or more best candidates

- *alpha* - Boolean (1 or 2 points awarded): belongs to several candidate groups: yes/no?
- *beta* - frequency in the union of candidate groups
- $z(W_b)$ - target function (alpha*beta)
- *gamma* - position in the sentence

Values of the Banker Joke

	alpha	*beta*	$z(W_B)$	*gamma*
be	1	1	1	4
use	2	1	2	9
lose	2	1	2	2
interest	2	2	**4**	**10**
banker	2	1	2	6

Precision:
$z(W_B)$ Sense-based method 0.2373
gamma Last word method 0.5145

PunFields at SemEval: Pun Detection

	homographic				heterographic			
system	**P**	**R**	**A**	**F_1**	**P**	**R**	**A**	**F_1**
Duluth	0.7832	0.8724	**0.7364**	**0.8254**	0.7399	0.8662	0.6871	0.7981
Idiom Savant	—	—	—	—	**0.8704**	0.8190	**0.7837**	**0.8439**
JU_CSE_NLP	0.7251	**0.9079**	0.6884	0.8063	0.7367	**0.9402**	0.7174	0.8261
PunFields	**0.7993**	0.7337	0.6782	0.7651	0.7580	0.5940	0.5747	0.6661
UWAV	0.6838	0.4723	0.4671	0.5587	0.6523	0.4178	0.4253	0.5094
random	0.7142	0.5000	0.5000	0.5882	0.7140	0.5000	0.5000	0.5882
ECNU*	0.7127	0.6474	0.5628	0.6785	0.7807	0.6761	0.6333	0.7247
Fermi†	**0.9024**	0.8970	**0.8533**	**0.8997**	—	—	—	—
N-Hance	0.7553	**0.9334**	**0.7364**	**0.8350**	0.7725	0.9300	0.7545	**0.8440**

PunFields at SemEval: Pun Location

	homographic				heterographic			
system	**P**	**R**	**A**	**F_1**	**P**	**R**	**A**	**F_1**
Duluth	0.7832	0.8724	**0.7364**	**0.8254**	0.7399	0.8662	0.6871	0.7981
Idiom Savant	—	—	—	—	**0.8704**	0.8190	**0.7837**	**0.8439**
JU_CSE_NLP	0.7251	**0.9079**	0.6884	0.8063	0.7367	**0.9402**	0.7174	0.8261
PunFields	**0.7993**	0.7337	0.6782	0.7651	0.7580	0.5940	0.5747	0.6661
UWAV	0.6838	0.4723	0.4671	0.5587	0.6523	0.4178	0.4253	0.5094
random	0.7142	0.5000	0.5000	0.5882	0.7140	0.5000	0.5000	0.5882
ECNU*	0.7127	0.6474	0.5628	0.6785	0.7807	0.6761	0.6333	0.7247
Fermi†	**0.9024**	0.8970	**0.8533**	**0.8997**	—	—	—	—
N-Hance	0.7553	**0.9334**	**0.7364**	**0.8350**	0.7725	0.9300	0.7545	**0.8440**