Computational Lexical Semantics: Methods and Applications

Alexander Panchenko
Language Technology Group, TU Darmstadt, Germany
panchenko@lt.informatik.tu-darmstadt.de

September 12, 2015

Plan

Plan

Motivation

Я обычно вечером пью **Ягуар** с пацанами на районе.

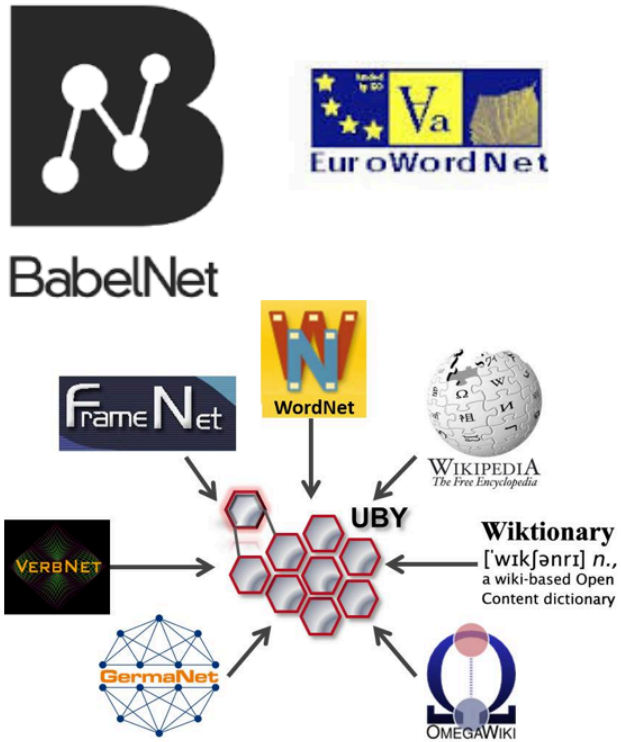
Отвертку
Водку
Пиво
Коньяк

Я наслаждаюсь вождением моего **Ягуара** 68-го года выпуска по воскресеньям.

Бентли
Мерседес
кабриолет
олд-таймер

Two Worlds of Computational Lexical Semantics

Human-Driven Approach



?

Data-Driven Approach

JoBimText
Linking Language to Knowledge with Distributional Semantics

word2vec
Tool for computing continuous distributed representations of words.

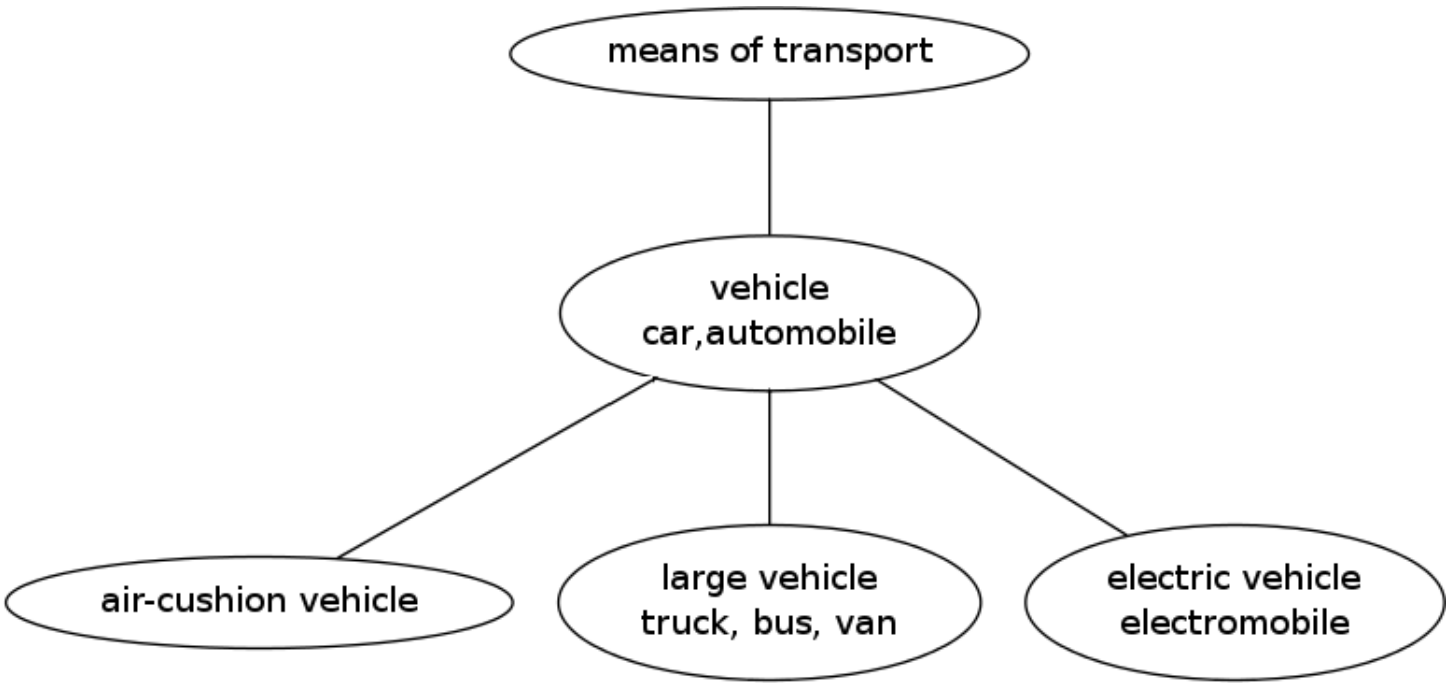
serelex

Semantic Vectors

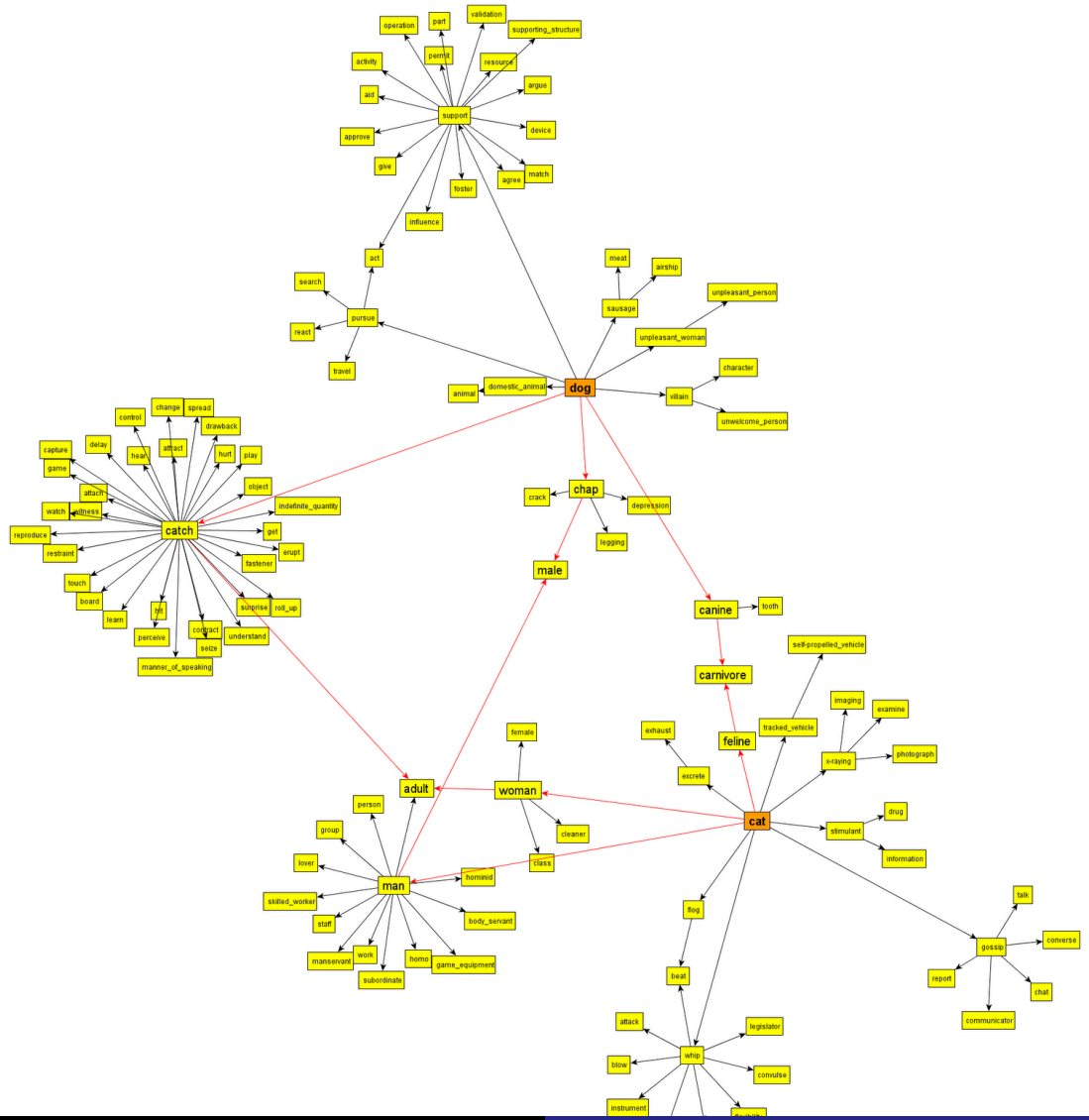
Semantic Resources

A **semantic resource** is an graph (C, R):

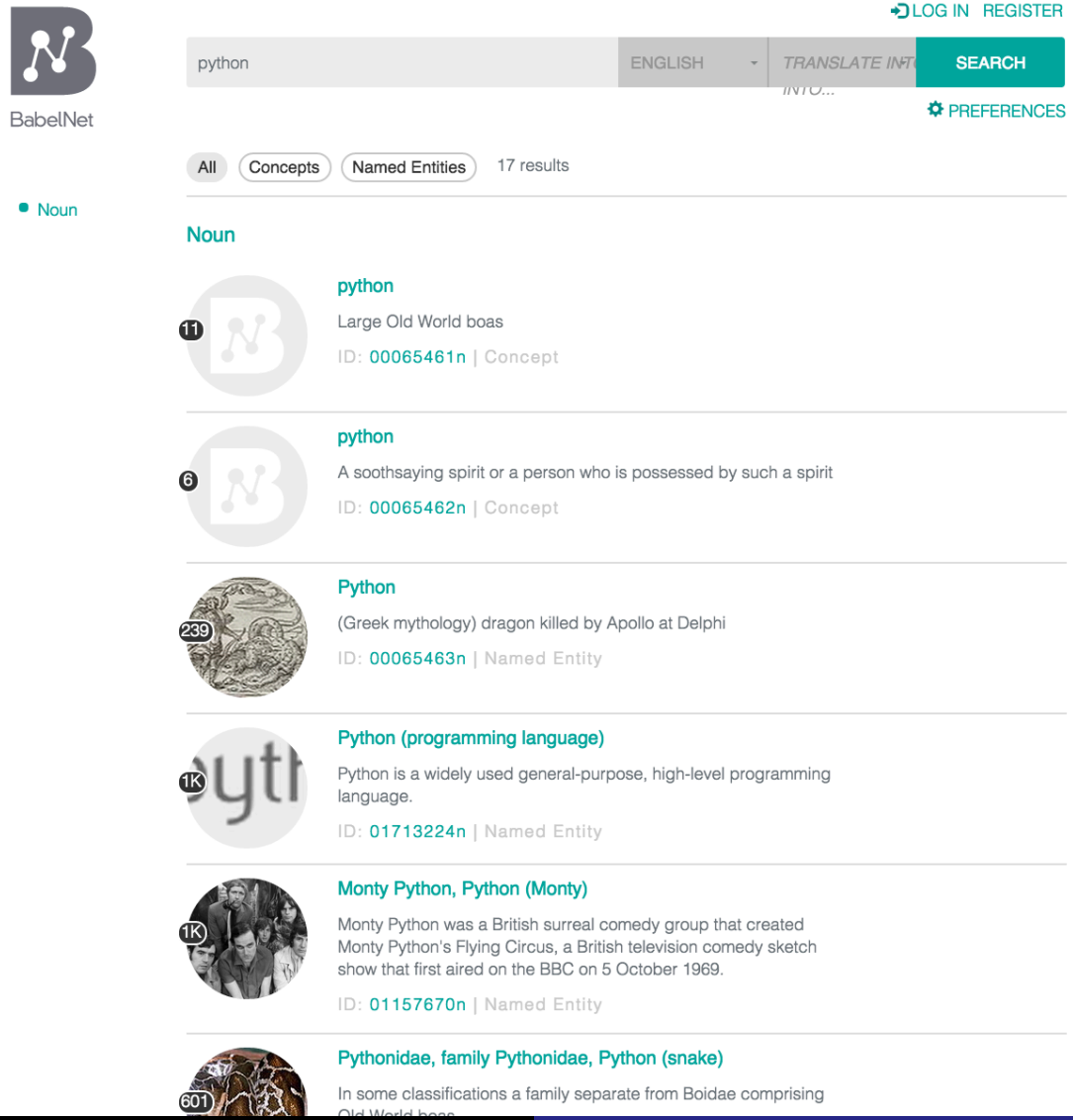
- nodes C represent **terms**;
- edges R represent untyped **semantic relations**.



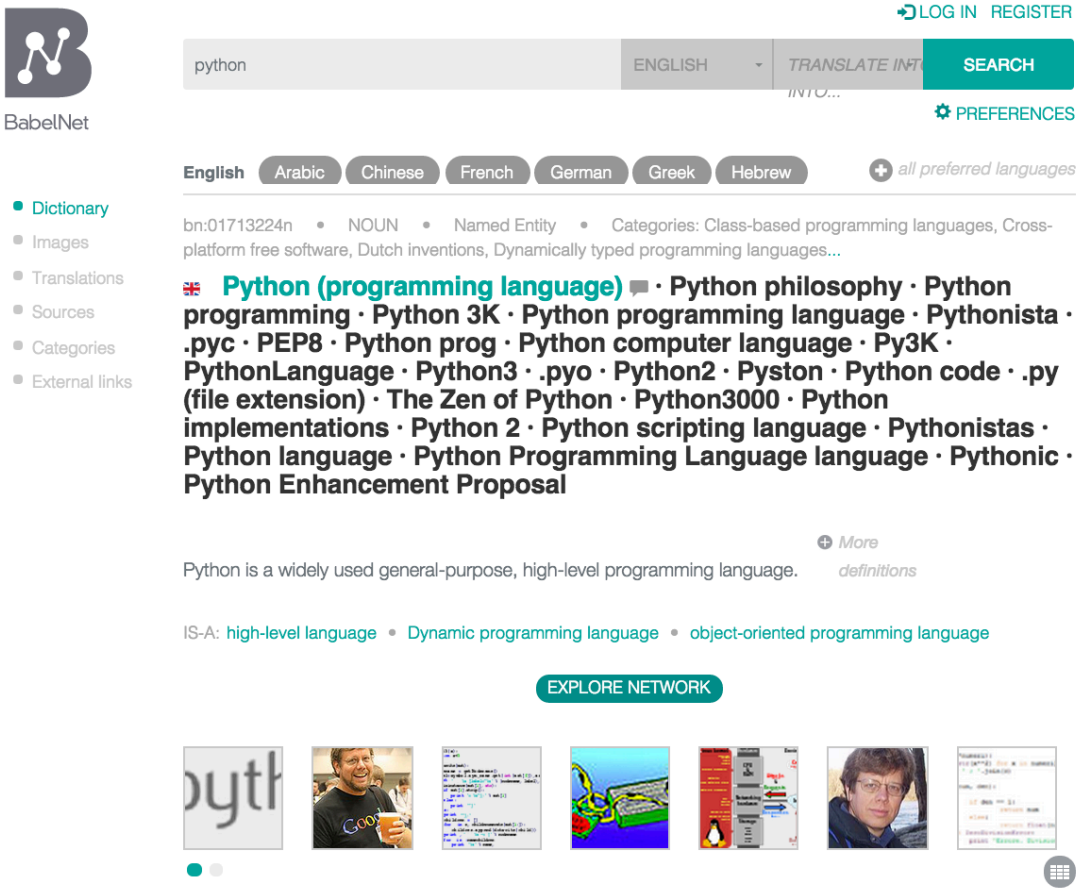
Semantic Resources: WordNet



A multilingual WordNet: BabelNet.org



A multilingual WordNet: BabelNet.org



Plan

Semantic Similarity Measures

Definition

A semantic similarity measure quantifies semantic relatedness input terms c_i, c_j with the similarity score $s_{ij} = sim(c_i, c_j)$:

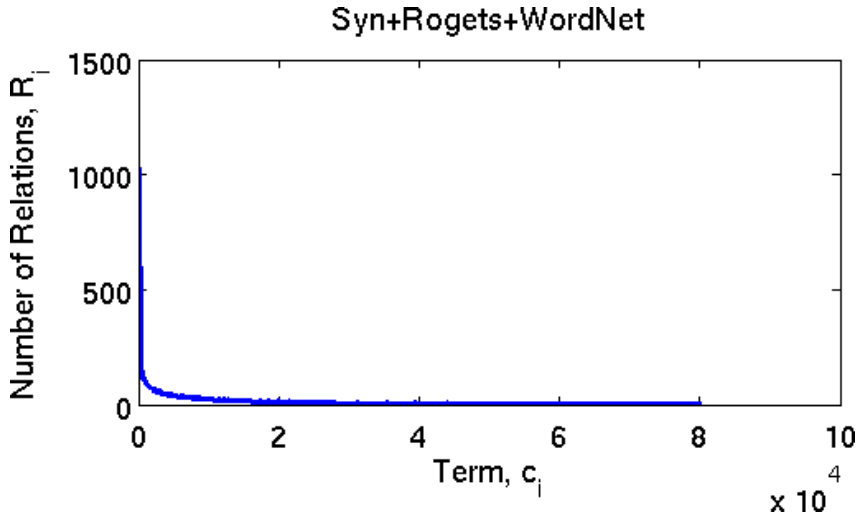
$$s_{ij} = \begin{cases} 1 & \text{if } \langle c_i, c_j \rangle \text{ is a pair of } syn, hyper, cohypo \\ 0 & \text{otherwise} \end{cases}$$

Properties

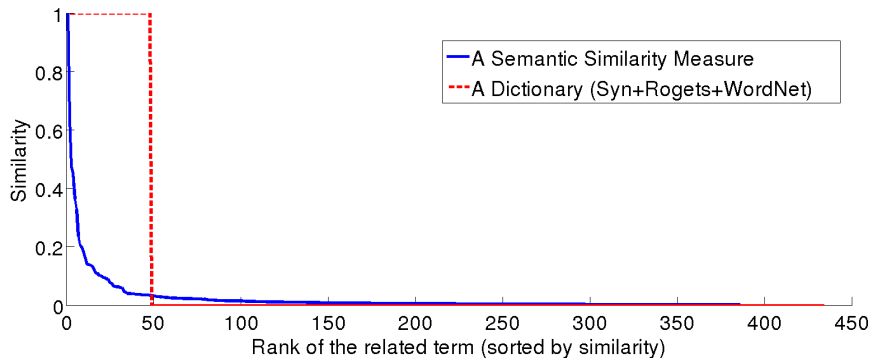
- Nonnegativity: $0 \leq s_{ij} \leq 1$;
- Reflexivity: $s_{ij} = 1 \Leftrightarrow c_i = c_j$;
- Symmetry: $s_{ij} = s_{ji}$;

Semantic Similarity Measures

- Many dissimilar pairs, few similar pairs: $s_{ij} \sim exp(\lambda)$:



- Similarity distribution of the term "doctor":



Plan

A live demo

- http://serelex.org

ferrari Search

Results count: 367

1. porsche
2. maserati
3. lamborghini
4. marque
5. bmw
6. audi
7. mercedes
8. jaguar
9. mclaren
10. dodge
11. bugatti
12. lancia
13. lotus
14. isuzu
15. tvr
16. vw
17. nissan
18. rally driving
19. fiat
20. mazda

Show all results...

computational linguistics Search

Results count: 88

1. psycholinguistics
2. machine learning
3. computer science
4. knowledge representation
5. cognitive science
6. artificial intelligence
7. information retrieval
8. neuroinformatics
9. natural language
10. graduate student
11. library science
12. distributed computing
13. research community
14. information processing
15. subfield
16. language acquisition
17. computer vision
18. soas
19. sociolinguistics
20. data mining

Show all results...

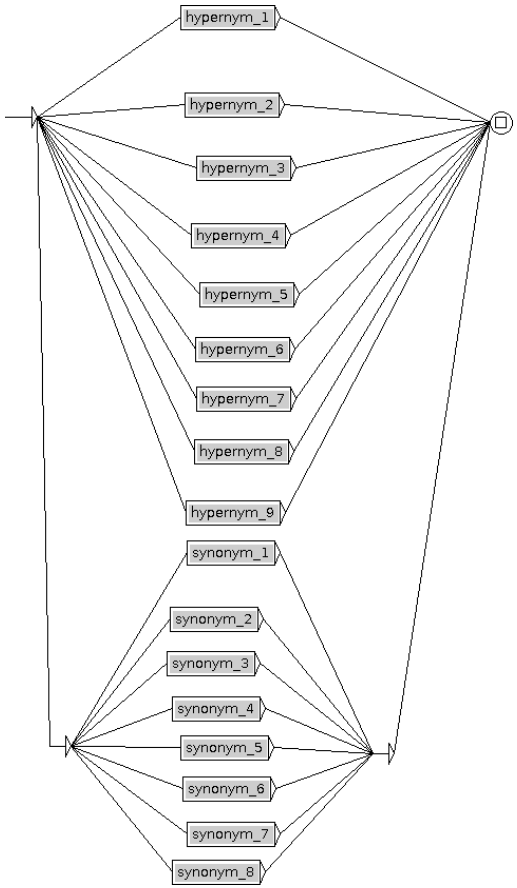
Lexico-syntactic patterns

- 18 patterns that extract **hypernyms**, **co-hyponyms** and **synonyms**

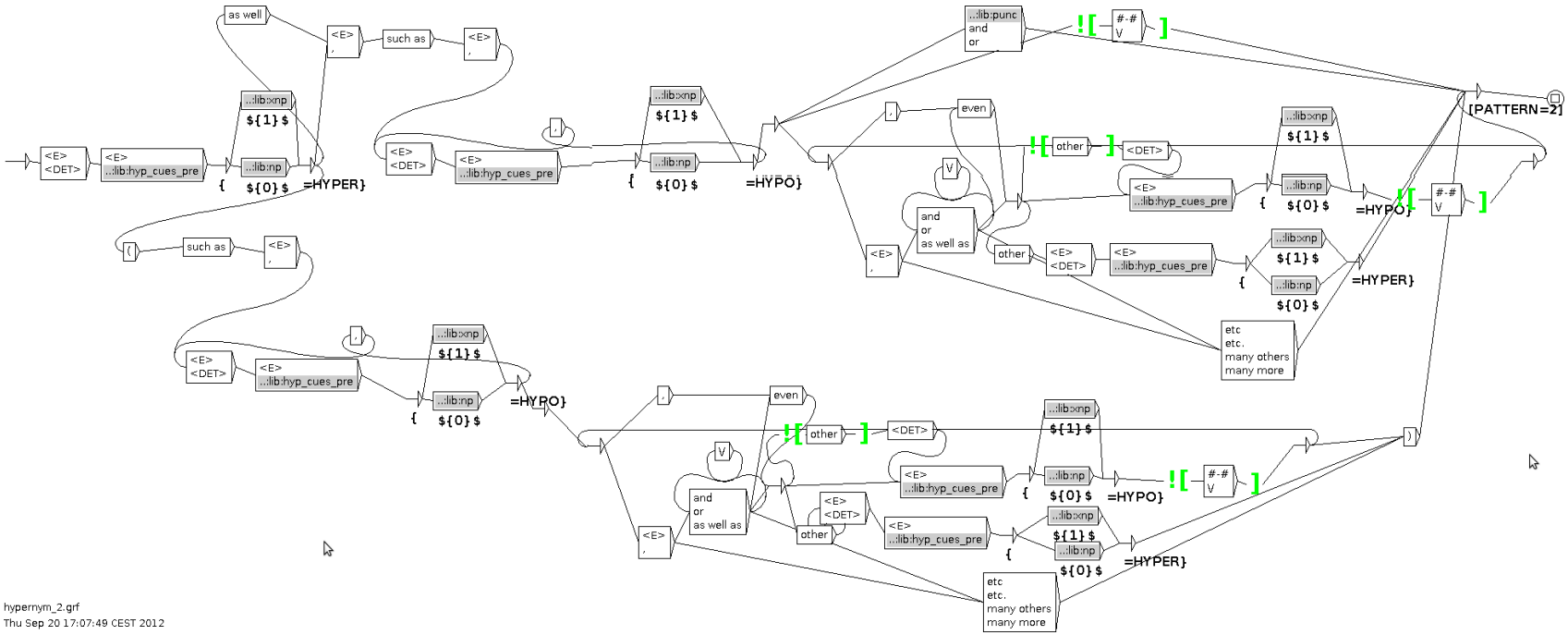
- ☐ *such NP as NP, NP[,] and/or NP;*
- ☐ *NP such as NP, NP[,] and/or NP;*
- ☐ *NP, NP [,] or other NP;*
- ☐ *NP, NP [,] and other NP;*
- ☐ *NP, including NP, NP [,] and/or NP;*
- ☐ *NP, especially NP, NP [,] and/or NP;*
- ☐ *NP: NP, [NP,] and/or NP;*
- ☐ *NP is DET ADJ.Superl NP;*
- ☐ *NP, e. g., NP, NP[,] and/or NP;*
- ☐ *NP, for example, NP, NP[,] and/or NP;*
- ☐ *NP, i. e.[,] NP;*
- ☐ *NP (or NP);*
- ☐ *NP means the same as NP;*
- ☐ *NP, in other words[,] NP;*
- ☐ *NP, also known as NP;*
- ☐ *NP, also called NP;*
- ☐ *NP alias NP;*
- ☐ *NP aka NP.*

Patterns are encoded as FSTs

- Finite State Transducers (FSTs)
- Open source corpus processing tool `Unitex`:
  `http://igm.univ-mlv.fr/~unitex/`



A pattern encoded as an FST



- Take into account linguistic variation
- Unlike string-based patterns (Bollegala et al., 2007)

Patterns extract concordances

- such diverse {[occupations]} as {[doctors]}, {[engineers]} and {[scientists]}[PATTERN=1]
- such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}[PATTERN=1]
- {traditional[food]}, such as {[sandwich]},{[burger]}, and {[fry]}[PATTERN=2]

Corpus

Corpus Wikipedia+ukWaC: $2.9 \cdot 10^{12}$ tokens

Name	# Documents	# Tokens	# Lemmas	Size
WaCky	2,694,815	$2,026 \cdot 10^9$	3,368,147	5.88 Gb
ukWaC	2,694,643	$0.889 \cdot 10^9$	5,469,313	11.76 Gb
WaCky + ukWaC	5,387,431	$2.915 \cdot 10^9$	7,585,989	17.64 Gb

Table 2.5: Corpora used by the PatternSim measure.

Extracted concordances

- Wikipedia – 1.196.468
- ukWaC – 2.227.025
- WaCypedia+ukWaC – 3.423.493

PatternSim Semantic Similarity

$$s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}.$$

- $P(c_i, c_j) = \frac{e_{ij}}{\sum_{ij} e_{ij}}$ – extraction probability of the pair $\langle c_i, c_j \rangle$, e_{ij} – frequency of co-occurrence of c_i and c_j in concordances K
- $P(c_i) = \frac{f_i}{\sum_i f_i}$ – probability of the term c_i, f_i – frequency of c_i
- $b_{i*} = \sum_{j:e_{ij} \geq \beta} 1$ – the number of extractions for term c_i with the frequency $\geq \beta$, $\mu_b = \frac{1}{|C|} \sum_{i=1}^{|C|} b_{i*}$ – the average number of extractions per term
- $p_{ij} \in [1; 18]$ – number of distinct patterns which extracted the relation $\langle c_i, c_j \rangle$

Semantic Relation Ranking

- Precision is **comparable or better** w.r.t. the baselines;
- Recall is **lower** w.r.t. the baselines.

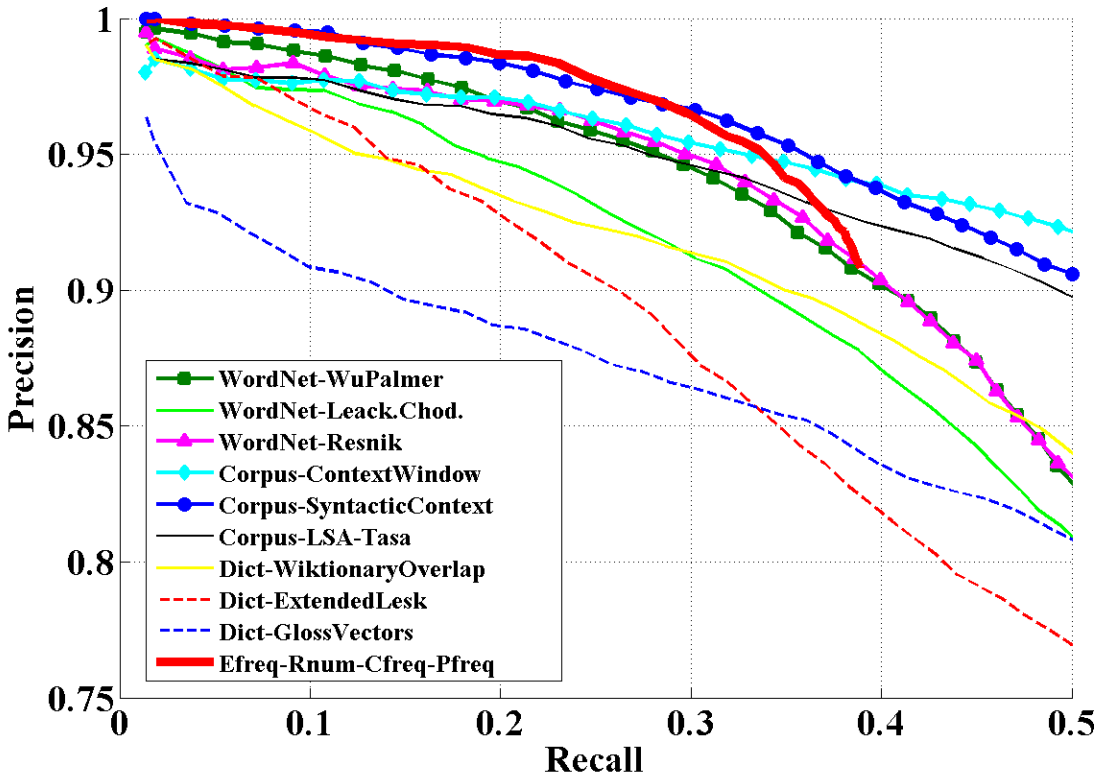
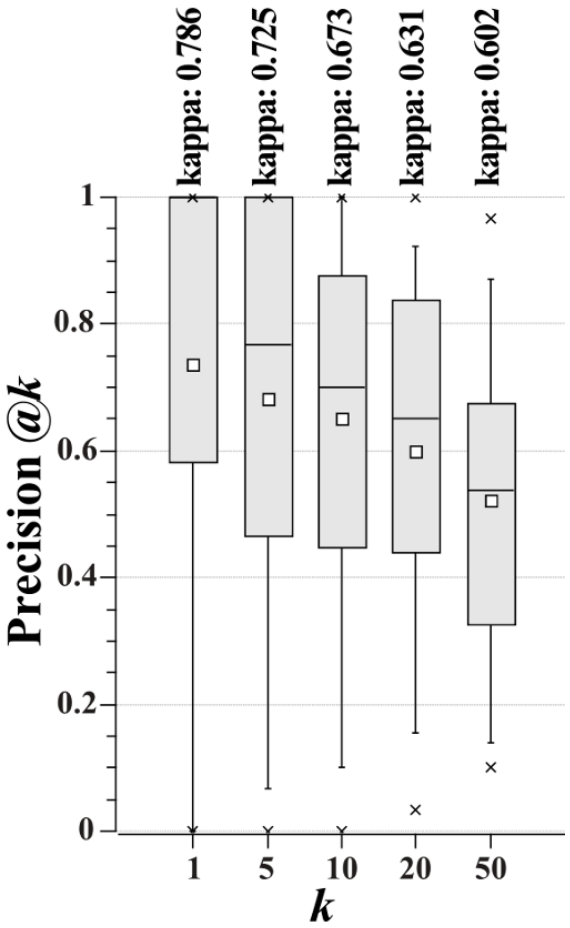


Figure: Precision-Recall graphs (the BLESS dataset).

Semantic Relation Extraction



- *Precision*@1 $\approx$ 0.80;
- "Good" coverage:

computational linguistics Search

System finds semantically related words.
For example, cottage cheese

Results count: 88

1 psycholinguistics
2 machine learning
3 computer science
4 knowledge representation
5 cognitive science
6 artificial intelligence
7 information retrieval
8 neuroinformatics
9 natural language
10 graduate student
11 library science

Word to search for: computational linguistics Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show W

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **computational linguistics** (the use of computers applications)
  - *direct hyponym* / *full hyponym*
    - S: (n) machine translation, MT (the use of co one language to another)
  - ***direct hypernym*** / *inherited hypernym* / *sister tern*
    - S: (n) linguistics (the scientific study of langua

Plan

Hybrid vs Single Measures

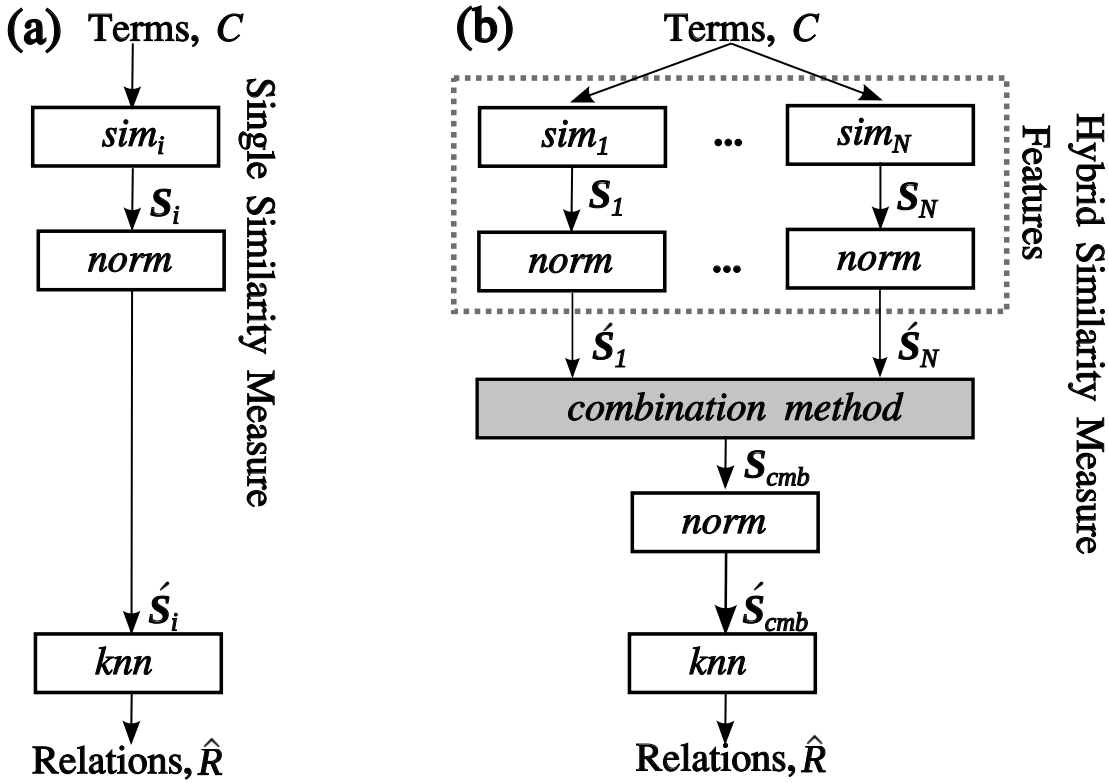


Figure: Semantic relation extractor based on:

- **(a)** a single similarity measure;
- **(b)** a hybrid similarity measure.

16 Features = 16 Single Similarity Measures

- 5 **network-based** measures :
  1. WuPalmer;
  2. Leacock and Chodorow;
  3. Resnik;
  4. Jiang and Conrath;
  5. Lin.
- 3 **web-based** measures (NGD-Yahoo/Bing/Google);
- 5 **corpus-based** measures:
  - 2 distributional (BDA, SDA)
  - 1 lexico-syntactic patterns (PatternSim)
  - 2 other co-occurence based (LSA, NGD-Factiva)
- 3 **definition-based** measures
  1. ExtendedLesk;
  2. GlossVectors;
  3. DefVectors-WktWiki.

Implementation of the baseline measures

- **Semantic Vectors:**
  `https://code.google.com/p/semanticvectors/`
- **S-Space Package:**
  `https://code.google.com/p/airhead-research/`
- **WordNet::Similarity:**
  `http://wn-similarity.sourceforge.net`
- **NLTK:** `http://nltk.googlecode.com/svn/trunk/doc/howto/wordnet.html`
- **WikiRelate!**
- **PatternSim:** `http://serelex.org`
- **Web-based metrics:**
  `http://cwl-projects.cogsci.rpi.edu/msr`
- **LSA:** `http://lsa.colorado.edu`

Supervised Combination of Measures

8 Logistic Regression

- A binary **logistic regression**;
- **Positive examples** – synonyms, hyponyms, co-hyponyms;
- **Negative examples** – random relations;
- A relation $\langle c_i, t, c_j \rangle \in R$ is represented with a **vector of pairwise similarities**: $\mathbf{x} = (s_{ij}^1, \ldots, s_{ij}^N), N = \overline{2, 16}$;
- Category y_{ij}:

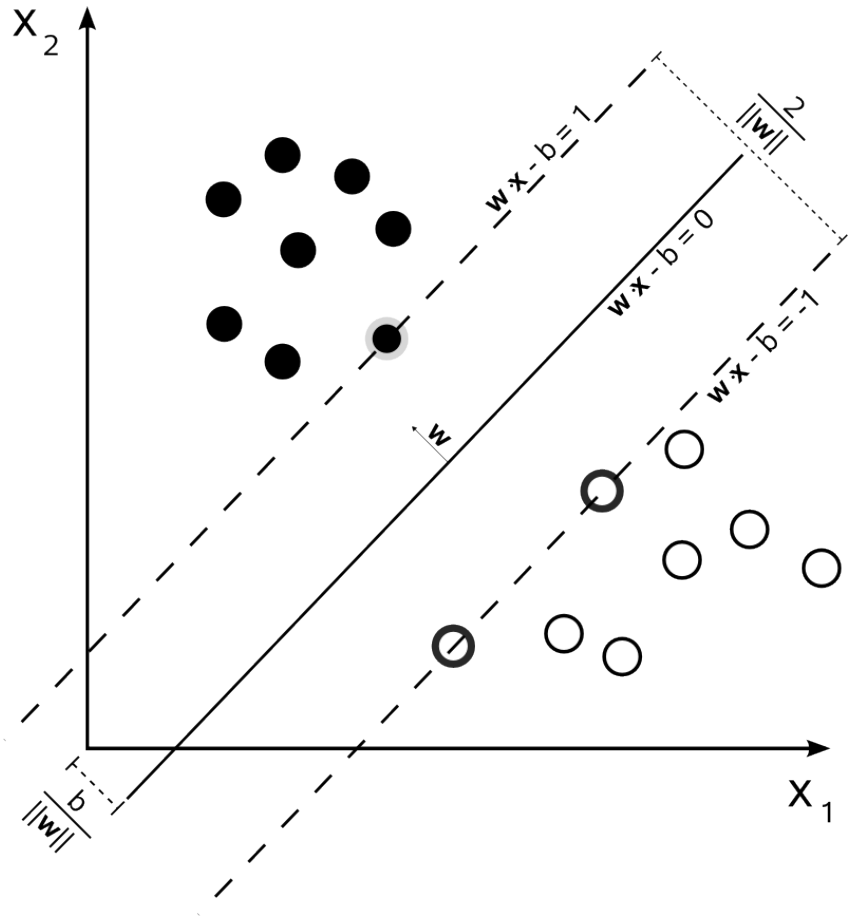
$$y_{ij} = \begin{cases} 0 & \text{if } \langle c_i, t, c_j \rangle \text{ is a random relation} \\ 1 & \text{otherwise} \end{cases}$$

- **Using the model** $(w_1, \ldots, w_K)$ for combination:

$$s_{ij}^{cmb} = \frac{1}{1 + e^{-z}}, z = \sum_{k=1}^{K} w_k s_{ij}^k + w_0.$$

Supervised Combination Methods

9 SVM.



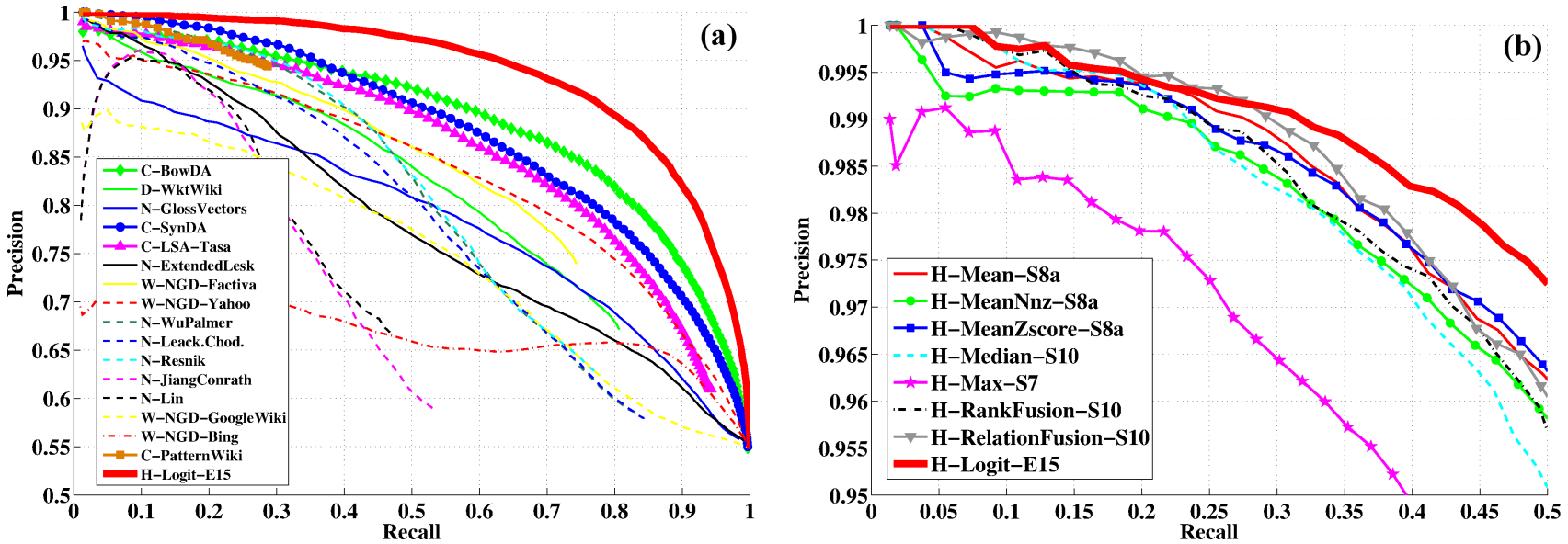
- The weights $\mathbf{w}$ and the support vectors SV:

$$\mathbf{w} = \sum_{x_i \in SV} \alpha_i y_i \mathbf{x}_i.$$

- **Using the model**

$$s_{ij}^{cmb} = \mathbf{w}^T \mathbf{x} + b = \sum_{k=1}^{K} w_i s_{ij}^k + b.$$

Hybrid Similarity Measures



Precision-Recall graphs calculated on the BLESS dataset:

- **(a)** 16 single measures and the best hybrid measure Logit-E15;
- **(b)** 8 hybrid measures.

Supervised Hybrid Similarity Measures

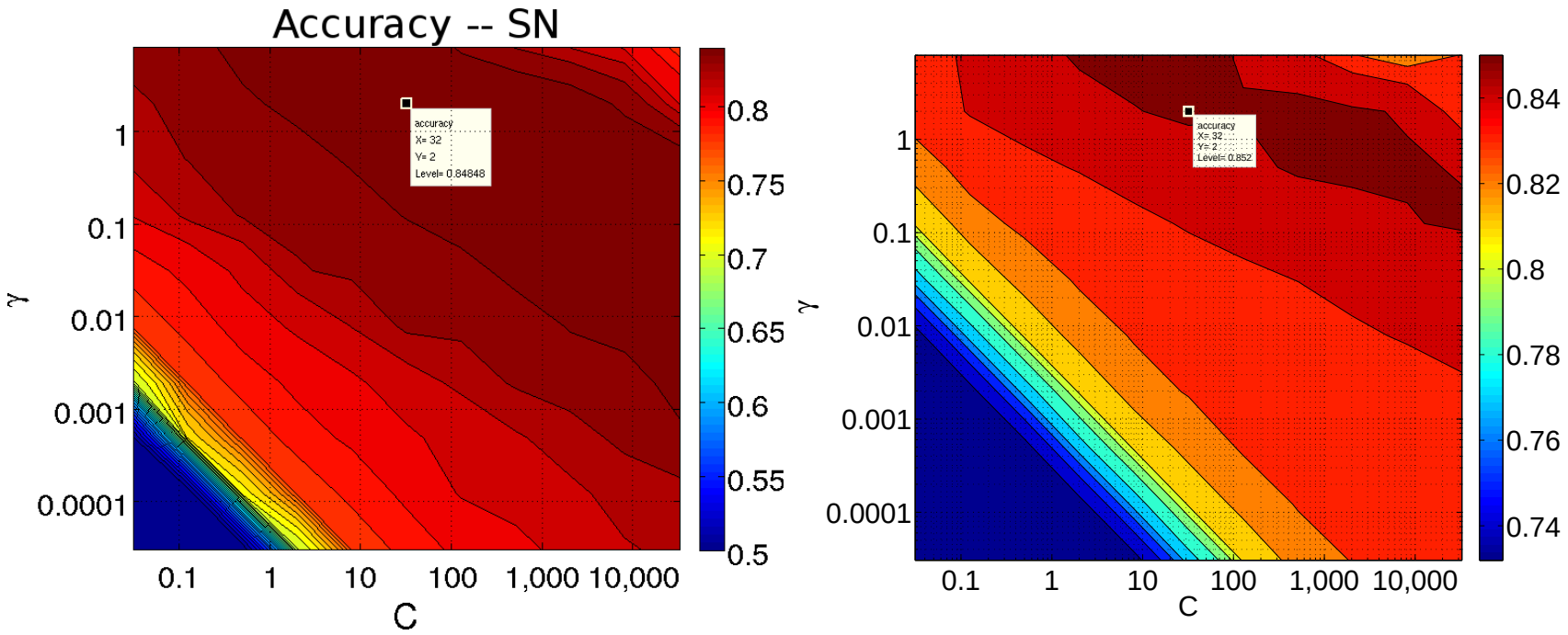


Figure: Meta-parameter optimization with the grid search of the C-SVM-radial-E15 measure.

Plan

Key facts

Word embedding is a dense vector representing a word obtained during a training of a neural network on a big corpus.

- Semantic similarity is cosine of word embeddings.
- The training process is known as representation learning.
- Learning methods rely on distributional hypothesis of Harris (1954), similarly to classical distributional models.
- Most popular representation learning methods:
  - Continuous Bag of Words (CBOW)
  - Skip-Gram Model
  - Global Vectors of for Word Representation (GloVe)

Distributional Hypothesis: "You shall know the word by the company it keeps" (Firth, 1957)

he curtains open and the stars shining in on the barely
ars and the cold , close stars " . And neither of the w
rough the night with the stars shining so brightly , it
made in the light of the stars . It all boils down , wr
surely under the bright stars , thrilled by ice-white
sun , the seasons of the stars ? Home , alone , Jay pla
m is dazzling snow , the stars have risen full and cold
un and the temple of the stars , driving out of the hug
in the dark and now the stars rise , full and amber a
bird on the shape of the stars over the trees in front
But I could n't see the stars or the moon , only the
they love the sun , the stars and the stars . None of
r the light of the shiny stars . The plash of flowing w
man 's first look at the stars ; various exhibits , aer
rief information on both stars and constellations, inc

Construct vector representations

	shining	bright	trees	dark	look
stars	38	45	2	27	12

Skip-Gram Model

- Probability that word w appears in some context c:

$$P(D = 1|w, c; \theta) = \frac{1}{1 + \exp^{-V_c W_w}}$$

- Assign high probability to (c, w) pairs which appear in texts (*corp*) and low probability to the ones which cannot (*rand*):

$$\theta^* = \textit{argmax} \prod_{(c,w) \in corp} P(D = 1|w, c; \theta) \prod_{(c,w) \in rand} (1 - P(D = 1|w, c; \theta))$$

- $\theta = (V, W)$ are two matrices with columns V_c and W_w contain vectors of *dim* dimensions of the context c and the word w.
- The "embedding" of the word w is the vector W_w.

Shared Task on Russian Semantic Similarity

- `www.dialog-21.ru/en/evaluation/2015/semantic`
- **Relatedness track**: Synonyms, Hypernyms, Human Judgements about Semantic Similarity
- **Associative track**: Free associations

Evaluation

word1	word2	sim
петух (cock)	петушок (cockerel)	0.952
побережье (coast)	берег (shore)	0.905
тип (type)	вид (kind)	0.852
миля (mile)	километр (kilometre)	0.792
чашка (cup)	посуда (tableware)	0.762
птица (bird)	петух (cock)	0.714
война (war)	войска (troops)	0.667
улица (street)	квартал (block)	0.667
...	...	...
доброволец (volunteer)	девиз (motto)	0.091
аккорд (chord)	улыбка (smile)	0.088
энергия (energy)	кризис (crisis)	0.083
бедствие (disaster)	площадь (area)	0.048
производство (production)	экипаж (crew)	0.048
мальчик (boy)	мудрец (sage)	0.042
прибыль (profit)	предупреждение (warning)	0.042
напиток (drink)	машина (car)	0.000
сахар (sugar)	подход (approach)	0.000
лес (forest)	погост (graveyard)	0.000
практика (practice)	учреждение (institution)	0.000

Best Systems

Model ID	HJ	RT-AVEP	AE-AVEP	AE2-AVEP	Method Description
5-ae-3	0.7071	0.9185	0.9550	0.9835	Word2vec (skip-gram, window size 10, 300d vectors) on ruwac + lib.ru + ru-wiki, bigrams on the same corpus, synonym database, prefix dictionary, orthographic similarity
5-rt-3	**0.7625**	**0.9228**	0.8887	**0.9749**	Word2vec (skip-gram, window size 10, 300d vectors) on ruwac + lib.ru + ru-wiki, synonym database, prefix dictionary, orthographic similarity
1-ae-1	0.6378	**0.9201**	**0.9277**	**0.9849**	Desicion trees based on n-grams (Wikipedia titles and search queries), morphological features and Word2Vec
15-rt-2	0.6537	0.9034	**0.9123**	0.9646	Word2vec trained on 150G of texts from lib.rus.ec (skip-gram, 500d vectors, window size 5, 3 iteration, min cnt 5)
16-ae-1	0.6395	0.8536	**0.9493**	0.9565	GloVe (100d vectors) on RuWac (lemmatized, normalized)
9-ae-9	**0.7187**	0.8839	0.8342	0.9517	Word2vec CBOW with window size 5 on Russian National Corpus, augmented with skip-gram model with context window size 20 on news corpus
17-rt-1	**0.7029**	0.8146	0.8945	0.9490	Distributional vector-based model, window size 5, trained on RUWAC and NRC, plmi-weighting
9-ae-6	**0.7044**	0.8625	0.8268	0.9649	Word2vec CBOW model with context window size 10 trained on web corpus
15-rt-1	0.6213	0.8472	**0.9120**	**0.9669**	Word2vec trained on 150G of texts from lib. rus.ec (skip-gram, 100d vectors, window size 10, 1 iteration, min cnt 100)
1-rt-3	0.4939	**0.9209**	0.8500	**0.9723**	Logistic regression trained on synonyms, hyponyms and hypernyms on word2vec features with AUC maximization
12-rt-3	0.4710	**0.9589**	0.5651	0.7756	Applying knowledge extracted from Wikipedia and Wiktionary for computing semantic relatedness

Skip-Gram on 150Gb of lib.rus.ec corpus

1	бренди	лафрейг шерри шивас коньяка мадеры лафроэйг содовой бакарди коньяк коньяком кофе пунша солодового бокал виски ликера полбокала кир
2	физика	физике эйнштейна электродинамику генетика радиофизика геометрия физики релятивистская математика начертательная наука астробиологи
3	линия	вертикальная полоска точка линиями инстербургская линию прочерченная конфигурация мовиуса полукружная позиция лотербура пролечь оч
4	марс	окололунной луну мимас тригоне нептун явин сатурн плутона марса секстиле марсе венере венера планет сатурна меркурий залбатану мерку
5	омар	халим джеззар хафиз аль халид абдаллах абд муса ахмад хусейн амир абулаббас омару ибн умар исмаил хасан мохаммед хамид мухаммед а
6	книга	страница книжища трилогия брошюра переиздавалась рукопись книге издана изданная многотомная сборник брошюрка пьеса книги тетрадочк
7	кредит	агропромбанке ссуда кредиту кредитов кредитом русфинанс кредиты безналу заемщик заем кредитам беспроцентный беспроцентным банково
8	выгода	убыток профит выгод профиту максимизируется услуга маржа инвестиция взаимовыгодность выгодой невыгода возмездная суммма барыш св
9	пол	линолеум ненатертый цементный паркет линолеумовый паркетный ковер наземь дионосопулос линолиум пола стукнувшуюся неоструганые по
10	еда	едой провизия закуска трапеза вкусная невкусна поесть жратва жрачка легкоусвояемая невкусной водка питье выпивка шамовка еде вкусно ж
11	авария	атомэнергетические энергосетях авиакатастрофа ошибка аварией неприятность диверсия денеболой москваленстрое авариях протечка ходы
12	страна	супердержава столица нация страной грузия родина мильдендо полуколониальная цивилизация империя россия греция стране планета сверх
13	юрист	консультант журналист бреслауэр гитлин присяжный консультирующий правовед прокурор геронтолог юристом политэкономист юристов повер
14	имущество	отказанное собственность отчуждаемое награбленное выморочное секвестрованное недвижимого залогодателя недвижимость собственность
15	попкорн	сперминт батончики мюслями гамбургеры пиццу гранолы гранолу киткат пепси арахисовым твикс арахисовое сэндвичи фростиз бургеры корно
16	владение	крещеною неотчуждаемое собственность бенефициарное владением югерами владеют пронию владения владению владею владении наслед
17	вещь	вещичку херь безделушка вещью штуку подробность ценная самая вещи небесполезная носильная своетака неизносимая вещица драгоценно
18	природа	безвременна сущность доразвилась дуальна нетварная человеческая недвойственна техносфера несотворенная расклассифицированная при
19	рождаемость	демографического брачности прироста деторождаемости нефтедобыча покупательная сверхсмертность репродуктивность смертности рождае
20	кровопролитие	грабежи сражение побоище братоубийственная пролитие усобной братоубийственное кровопролития предотвратить бесчинство междуусобиц
21	епископ	епископа поппон лангрский архидиакон митрополит монтейский священник хинкмар прелат настоятель бовэзский нарбонский бовезский архие
22	доктор	шатонней глайстер доктора лагенторп стифен врача чакк нейропсихиатр мистер сайболт доктором невролог бичэма доктору мейенстейн обер
23	изображение	увеличенное объемное изображением телеизображение изображения изображению голографическое очертание спроецировалось видеоизобр
24	огурец	цукини пошинкуйте салат огурцом репчатая накрошенный стручковый баклажан лимон полпучка помидоры шинкуется помидорчик огурца арбу
25	стекло	запотевающее витринное стеклом смотровое разбитое ветрозащитное бемское оконное полуопущенное стекле стекла незатонированное бэмс
26	улика	кастете компра зацепка штульцевскую килбэрна зацепочка заказуха мэллинса дактилограф скитальному куплетовские платежка улике уликах
27	отмывание	отмывки обналичка отмывка отмывке букмекерство наркобизнес невозврат офшорные зарабатывание мошенничество крышевание отмывани
28	страх	снедавший ненависть панический обуял побарывает закравшийся атавистичное овладевавший трепет оцепеняющий всепоглощающий обесси
29	среда	социосистема абиотическую высококонкурентная средой суперблагоприятная сложноорганизованная реагенная макроокружение биосистема
30	сезон	фестиваль дект ташриту сезоне сезона муссонов мусонных июль дождей муссонных триместр сноубордный осень мягмар лета летом несезон

Plan

Plan

Search for Related Words: the List and the Graph

- http://serelex.org

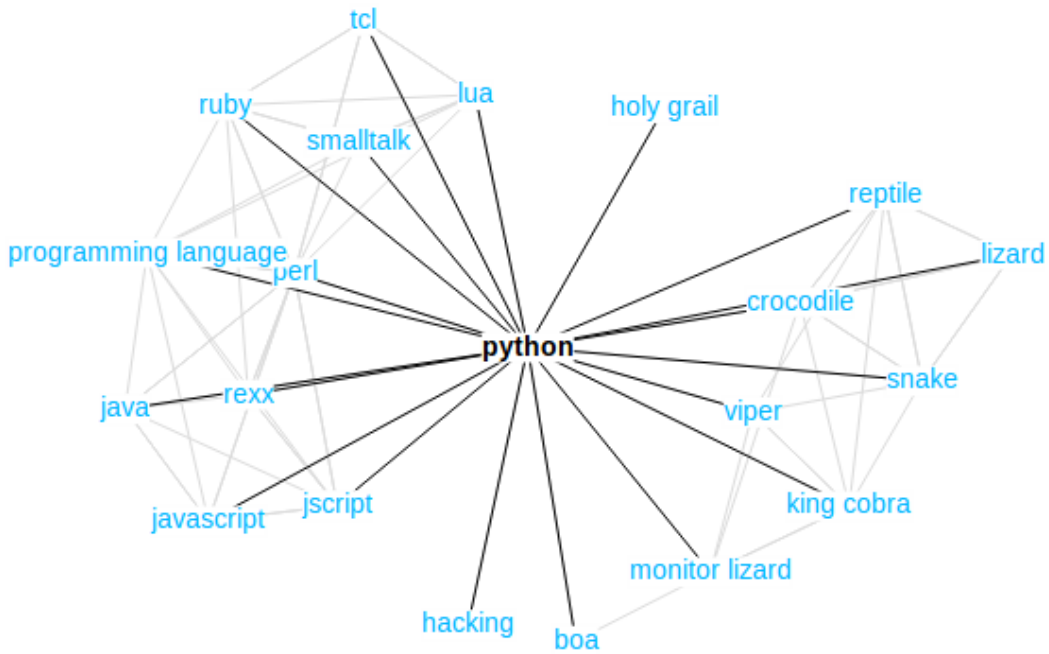
python Search

System finds semantically related words.
For example, lamborghini

Results count: 412

1 perl
2 ruby
3 programming language
4 tcl
5 monitor lizard
6 lua
7 boa
8 crocodile
9 holy grail
10 hacking
11 java
12 king cobra
13 reptile
14 snake
15 jscript
16 javascript
17 viper
18 smalltalk
19 rexx
20 lizard

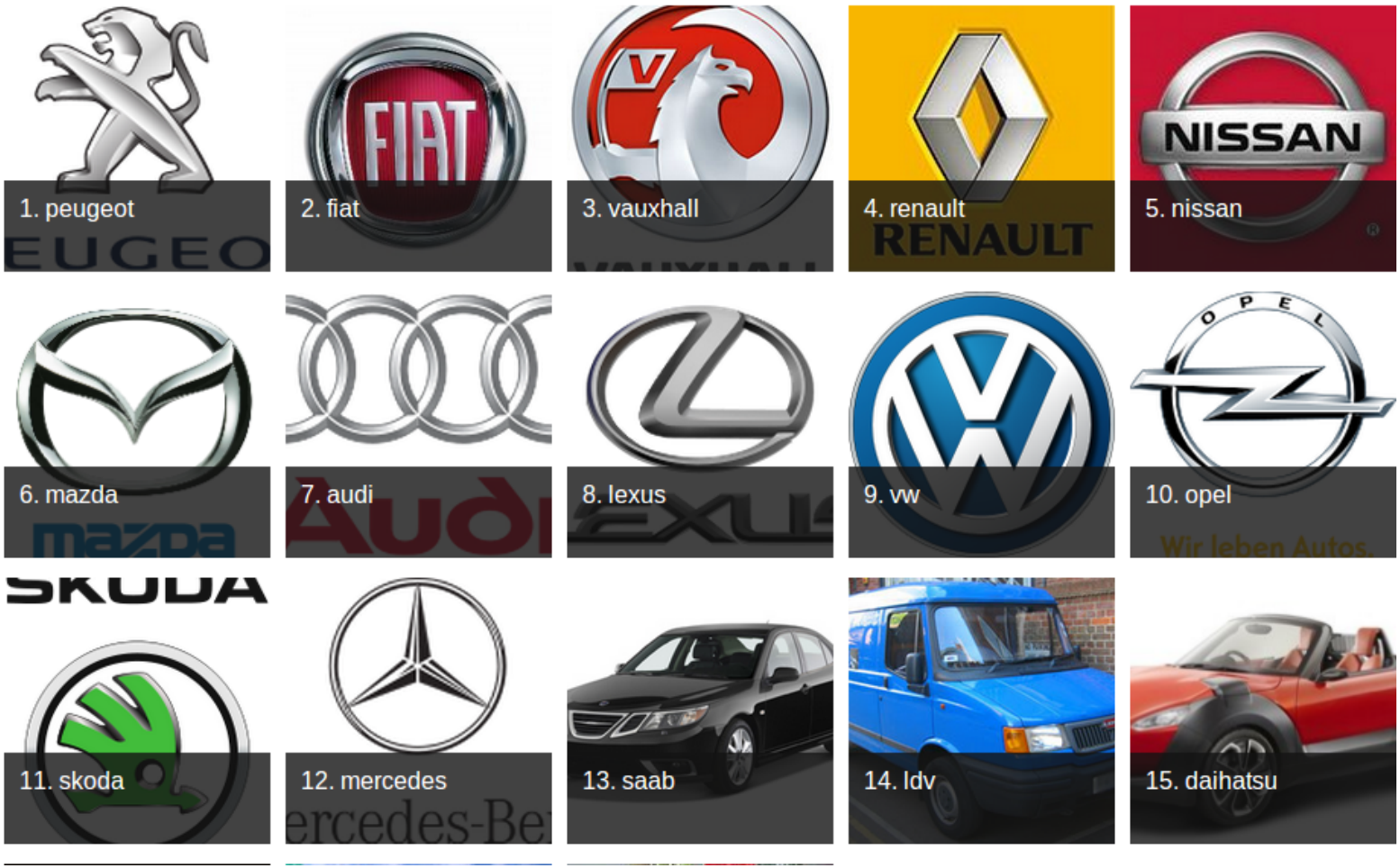
Show all results...



Search for Related Words: the Images

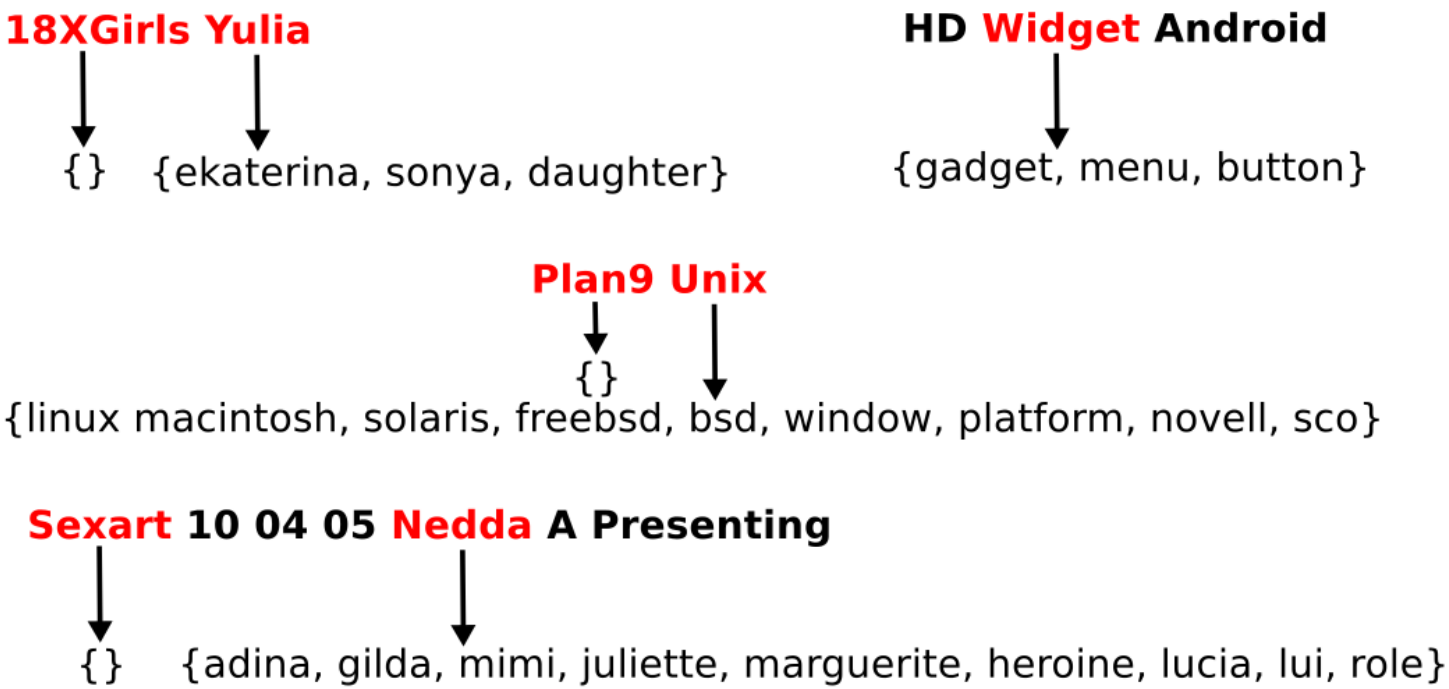
citroen Search

System finds semantically related words.
For example, linux



Plan

Short text classification with Vocabulary Projection



Evaluation of the Vocabulary Projection

Training Dataset	Test Dataset	Accuracy	Accuracy (voc. projection)
Gallery (train)	Gallery	96.41	**96.83** (+0.42)
PirateBay Title+Desc+Tags	PirateBay Title+Desc+Tags	**98.92**	98.86 (–0.06)
PirateBay Title+Tags	PirateBay Title+Tags	**97.73**	97.63 (–0.10)
Gallery	PirateBay Title+Desc+Tags	90.57	**91.48** (+0.91)
Gallery	PirateBay Title+Tags	84.23	**88.89** (+4.66)
PirateBay Title+Desc+Tags	Gallery	88.83	**89.04** (+0.21)
PirateBay Title+Tags	Gallery	91.16	**91.30** (+0.14)

Table: Performance of an C-SVM linear classifier (10-fold cross validation).