

Cross lingual and cross cultural textual encoding of opinions and sentiments

Dan Tufiş, Radu Ion



*Research Institute for Artificial Intelligence
Romanian Academy*

Why so much interest in subjectivity analysis ?

- Social community-oriented websites and user generated content are becoming an extremely valuable source of information for everyday information consumer. But also for various kinds of decision makers;
- Two main types of textual information on the web:
facts (*objective*) and opinions (*subjective*)
- Current search engines search for facts not opinions (current search ranking strategy is not appropriate for opinion search/retrieval)

- Word-of mouth on the web is sometimes perceived as being more trustful than the regular mass-media sources!
 - In user generated content (review sites, forums, discussion groups, blogs etc) one can find descriptions of personal experiences and opinions on almost anything;
 - valuable for common people for practical daily decisions (buying products or services, going to a movie/show, traveling somewhere, finding opinions on political topics or on various events...
 - valuable for professional decision makers in many areas, so they support this trend.

Who are the main interested parties?

- Among others:
 - Business and marketing intelligence analysts (product and services reviews, customers relationship management)
 - Military intelligence analysts (who feels like doing nasty things)
 - Political intelligence analysts (what's the attitude of voters towards specific political decisions, political candidates)
 - Artificial Intelligence analysts (doing research and implementing systems to support the above mentioned analysts)

Opinion mining

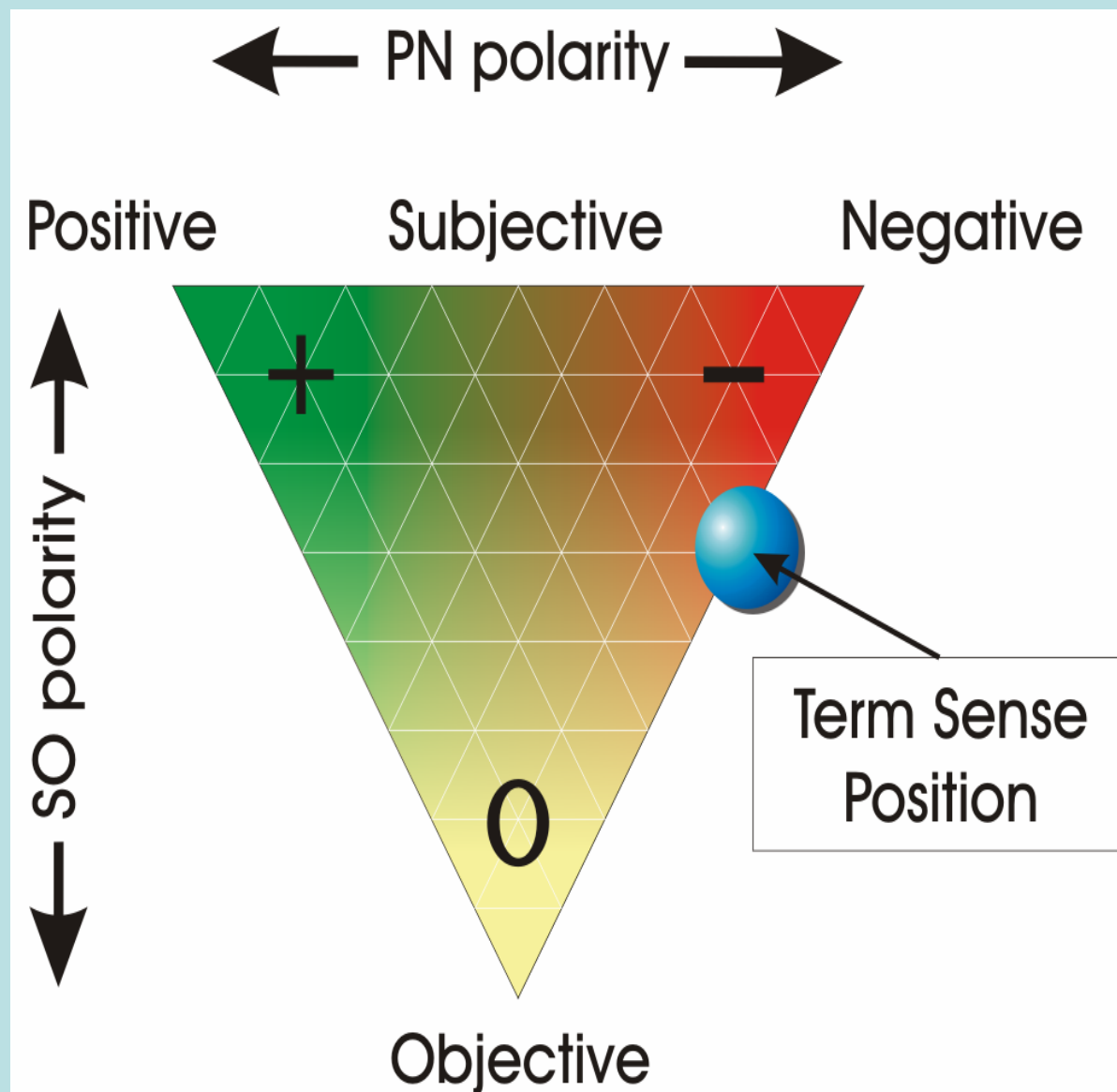
1. Identify opinionated sentences relevant for a specific topic
 - Hard, implemented as two steps procedure (identify opinionated sentences, and then judge their relevance to the topic)
2. Identify the opinion holder (person, institution, government, etc)
 - Hard, frequently requires anaphora resolution
3. Identify the attitude of the opinion holder (polarity of the opinion: positive, negative, neutral)
 - Hard, needs WSD, reasoning
4. Summary of the opinions and/or graphical visualisation of the results
 - Not as hard as 1-3, but far from being trivial

SentiWordNet

Andrea Esuli, Fabrizio Sebastiani. SentiWordNet: A publicly Available Lexical Resourced for Opinion Mining, LREC2006

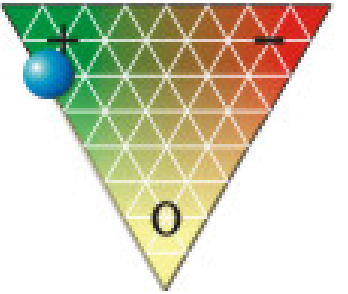
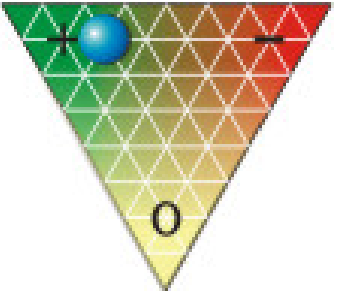
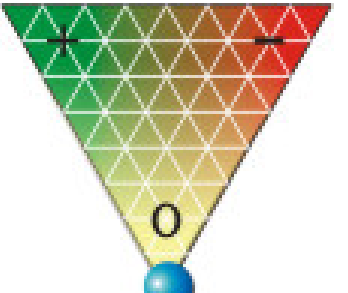
The basic assumptions:

1. words have graded polarities along the orthogonal axes:
Subjective-Objective (SO)
& **Positive-Negative (PN)**
1. The SO and PN polarities depend on the various senses of a given word (context)



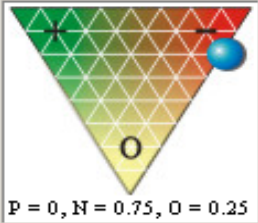
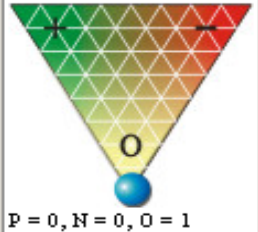
Adjective

3 senses found.

 <p>A ternary diagram with vertices P (top), N (bottom-left), and O (bottom-right). The diagram is divided into a 4x4 grid of smaller triangles. A blue sphere is positioned on the left edge (P-N) at the 3/4 mark from P. The diagram is shaded with a gradient from green at the top to red at the bottom.</p> <p>$P = 0.75, N = 0, O = 0.25$</p>	<p>estimable(1) <i>deserving of respect or high regard</i></p>
 <p>A ternary diagram with vertices P (top), N (bottom-left), and O (bottom-right). The diagram is divided into a 4x4 grid of smaller triangles. A blue sphere is positioned on the left edge (P-N) at the 1/4 mark from P. The diagram is shaded with a gradient from green at the top to red at the bottom.</p> <p>$P = 0.625, N = 0.25, O = 0.125$</p>	<p>honorable(5) good(4) respectable(2) estimable(2) <i>deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</i></p>
 <p>A ternary diagram with vertices P (top), N (bottom-left), and O (bottom-right). The diagram is divided into a 4x4 grid of smaller triangles. A blue sphere is positioned at the bottom vertex O. The diagram is shaded with a gradient from green at the top to red at the bottom.</p> <p>$P = 0, N = 0, O = 1$</p>	<p>computable(1) estimable(3) <i>may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</i></p>

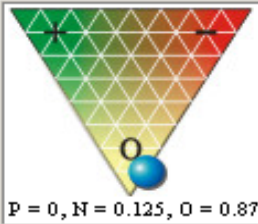
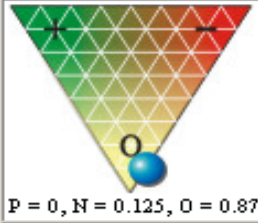
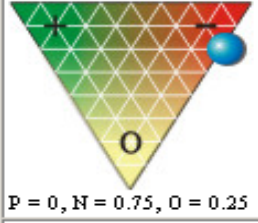
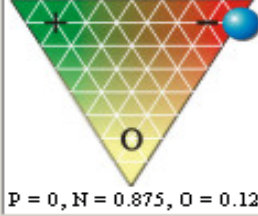
Verb

2 senses found.

	<p>short(1) short-change(1) <i>cheat someone by not returning him enough money</i></p>
	<p>short-circuit(2) short(2) <i>create a short-circuit in</i></p>

Adjective

15 senses found.

	<p>short(1) <i>primarily temporal sense; indicating or being c seeming to be limited in duration; "a short life" "a short flight"; "a short holiday"; "a short story"; "only a few short months"</i></p>
	<p>short(2) <i>primarily spatial sense; having little length or lacking in length; "short skirts"; "short hair"; "the board was a foot short"; "a short toss"</i></p>
	<p>short(3) <i>low in stature; not tall; "his was short and stocky"; "short in stature"; "a short smokestack"</i></p>
	<p>inadequate(2) poor(7) short(4) <i>not sufficient to meet a need; "an inadequate income"; "a poor salary"; "money is short"; "on short rations"; "food is in short supply"; "short on experience"</i></p>

State of the art

- Monolingual research: more and more numerous and in more and more languages
- Multilingual comparative studies (different comparable text-data, different languages): not very many, but their number is increasing
- We are not aware of cross-lingual studies (parallel texts)
 - why? Possible answers:
 - The original opinions are those expressed in the source language; the target language contains (presumably faithful) translations of the holders' opinions;
 - Expressing opinions is a cultural matter: most translations are concerned with the factual content preservation

Questions (Case 1)

- Assume a collection of original documents in Japanese S_{JP} and two translations of it in English T_{EN1} and T_{EN2} with the first translation done by a Japanese with a perfect command of English and the second translation done by an American with a perfect command of Japanese.
 - Would opinions in S_{JP} and T_{EN1} be “the same”?
 - Would opinions in S_{JP} and T_{EN2} be “the same”?
 - Would opinions in T_{EN1} and T_{EN2} be “the same”?“the same”=#opinionated sentences, polarity

Answers: No idea! Possible guesses: Yes?, yes?, YES?₁₀

Questions (Case 2)

- Assume a collection of original documents in Japanese S_{JP} containing reports (newspaper articles, news agencies briefs, official statements) on specific international events and a collection of documents in English S_{EN} containing reports of similar lengths and from corresponding sources on the same international events
 - Would opinions in S_{JP} and S_{EN} be “the same”?
“the same”=#opinionated sentences, polarity

Answer: Probably, no!

Why? Due to cultural differences. For instance, (cf Kim & Myaeng, NTCIR 2007) “a sentence in Japanese, reporting on a merge of two companies should be judged to have negative sentiment whereas the same kind of activities in the US would be a positive event”

Opinion analysis across languages NTCIR6

- David Kirk Evans, Lun-Wei Ku, Yohei Seki, Hsin-His Chen, Noriko Kando (2007)
 - Case 1.5 experiments (comparable texts) in Japanese, English and Chinese (English translations probably done by Japanese and Chinese employees of the local news agencies)
 - Japanese data(1998-99): Mainichi Daily News, Yomiuri
 - English data(1998-99): Mainichi Daily News, Korea Times, Xinhua
 - Chinese data(1998-99): United Daily News, China Times, China Time Express, China Daily News, etc

Language	Topics	Documents	Sentences	Opinionated lenient/strict	Relevant lenient/strict
Chinese	32	843	8546	62% / 25%	39% / 16%
English	28	439	8528	30% / 7%	69% / 37%
Japanese	30	490	12525	29% / 22%	64% / 46%

What does this experiment show?

- Big differences across languages in the Gold Standards
- Despite using similar approaches, big differences in the performances of the competing systems with respect to the processed language (best in Chinese, worst in English)
- Are these differences explained by the existing differences in the annotation?
 - Partly!
 - Annotators training could be a better explanation (big differences between the annotators in the three languages)
 - Language and cultural differences probably significantly mattered

Feature-based opinion and sentiment analysis

- The building blocks: *sentiment words*
- The words become sentiment words *in context*
- The *bag-of-words* approach works pretty bad (but works!) and there are various ways (maybe expensive) to improve the opinion and sentiment analysis
- Syntax and punctuation (usually discarded) play also an important role in judging the subjectivity of a piece of text.

Word-Senses and Subjectivity

- SentiWordNet associates subjectivity scores (P, R, O) to WordNet synsets, i.e. to the word-senses.
- Lexical semantics is very important here.
- WSD would be highly instrumental (e.g. JW&RM)
- Dependency Linking (which is less than parsing, but easier to obtain) is much more appropriate than B-o-W

Verb

2 senses found.

<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the upper right quadrant, indicating a high positive subjectivity score. The plot is labeled with P=0, N=0.75, O=0.25.</p>	<p>short(1) short-change(1) <i>cheat someone by not returning him enough money</i></p>
<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the lower left quadrant, indicating a high negative subjectivity score. The plot is labeled with P=0, N=0, O=1.</p>	<p>short-circuit(2) short(2) <i>create a short-circuit in</i></p>

Adjective

15 senses found.

<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the lower right quadrant, indicating a high positive subjectivity score. The plot is labeled with P=0, N=0.125, O=0.875.</p>	<p>short(1) <i>primarily temporal sense; indicating or being or seeming to be limited in duration; "a short life"; "a short flight"; "a short holiday"; "a short story"; "only a few short months"</i></p>
<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the lower right quadrant, indicating a high positive subjectivity score. The plot is labeled with P=0, N=0.125, O=0.875.</p>	<p>short(2) <i>primarily spatial sense; having little length or lacking in length; "short skirts"; "short hair"; "the board was a foot short"; "a short toss"</i></p>
<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the upper right quadrant, indicating a high positive subjectivity score. The plot is labeled with P=0, N=0.75, O=0.25.</p>	<p>short(3) <i>low in stature; not tall; "his was short and stocky"; "short in stature"; "a short smokestack"</i></p>
<p>A triangular subjectivity plot with a color gradient from green (top) to red (bottom). A blue dot is positioned in the upper right quadrant, indicating a high positive subjectivity score. The plot is labeled with P=0, N=0.875, O=0.125.</p>	<p>inadequate(2) poor(7) short(4) <i>not sufficient to meet a need; "an inadequate income"; "a poor salary"; "money is short"; "on short rations"; "food is in short supply"; "short on experience"</i></p>

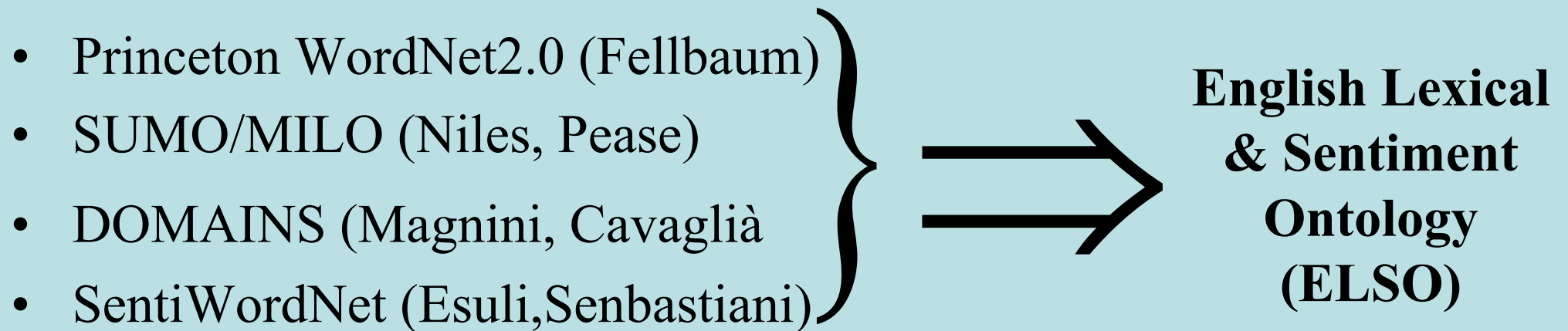
Cross-lingual opinion and sentiment analysis

- A parallel text (EN-XX) e.g, Orwell's "1984", MultiSemCor, Euro-Parl, JRC-Acquis etc.
- Word Align and WSD the EN-XX bitext
- Use a scoring method for the senti-words and valency-shifter words in each part of the bitext (based on SentiWordNet scores) to classify the opinionated sentences
- Try to answer Question 1
 - Evaluate monolingually (both in En and XX) whether the mark-ups hold true; for En you might use OpinionFinder (Wiebe, Riloff et al.) and compare its classification with the SentiWordNet-based classification
 - Write immediately a breakthrough paper (whatever the results of the evaluation)

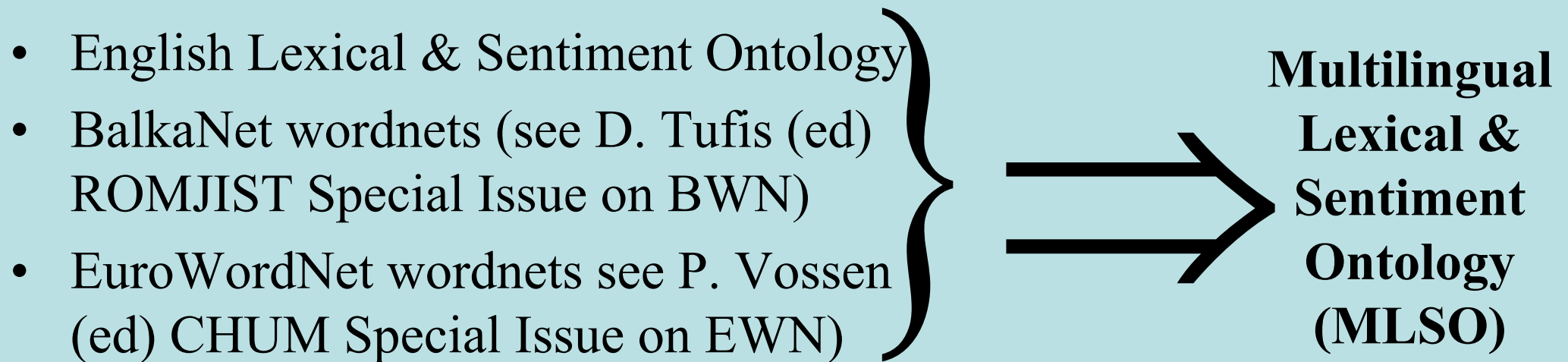
What you would need to do it?

- Quality multilingual lexical and sentiment marked-up resources (multilingual lexical and sentiment ontologies are probably the best)
- List of valency shifters and rules to define their scope and results on the sentiment words (Polanyi& Zaenen, 2006)
- Preprocessing tools (sentence alignment, tokenizer, POS taggers, lemmatizers, dependency linkers)
- Alignment tools (e.g. COWAL)
- WSD tools (WSD Tool, SynWSD)
- Sentence opinion scorer and classifier
- Annotation transfer tools

English Lexical and Sentiment Ontology (ELSO)



Multilingual Lexical and Sentiment Ontologies (MLSO)



The EuroWordNet and BalkaNet multilingual wordnets use the Princeton WordNet as the InterLingual Index (ILI) => any sentiment and ontological mark-up in PWN is available in the aligned monolingual wordnets; altogether they make a MLSO.

Some Quantitative Data about RoWordNet

- Synsets: 42,463
- Relations: 55,178
- Literals: 65,270
- SUMO/MILO labels: 38,538 (1821 concepts)
- DOMAINS labels: 49,563 (165 domains)
- Sentiment labeled synsets: 42,463

The encoding of a (sentiment) synset in RoWordNet

<SYNSET>

<ID>ENG20-04854135-n</ID> <BCS>3</BCS>

<DOMAIN>factotum</DOMAIN>

<SUMO>SubjectiveAssessmentAttribute<TYPE>+</TYPE></SUMO>

<POS>n</POS>

<SYNONYM><LITERAL>bine<SENSE>16</SENSE></LITERAL>

<LITERAL>bun<SENSE>51</SENSE></LITERAL>

<LITERAL>virtute<SENSE>2</SENSE></LITERAL>

</SYNONYM>

<DEF> Înclinație statornică specială către un anumit fel de îndeletniciri sau acțiuni frumoase. </DEF>

<ILR>ENG20-04521520-n<TYPE>hypernym</TYPE></ILR>

<ILR>ENG20-04855887-n<TYPE>near_antonym</TYPE></ILR>

<SENTIWN>

<P>0.75</P><N>0</N><O>0.25</O>

</SENTIWN>

</SYNSET>

animal

*[n] animal:1, creatură:1, jivină:1, lighioană:1, necuvântător:1

+[a] animal:3.1

[n] brută:1, bestie:50, animal:2

View Tree RevTree BCS1.2 BCS3 Edit Words ROM XML

POS: n ID: ENG20-00012748-n BCS: 1

Synonyms: animal:1, creatură:1, jivină:1, lighioană:1, necuvântător:1

Definition: Ființă organizată, uni-sau pluricelulară, înzestrată cu facultatea de a simți și de a se mișca

Domain: animals

Domain: biology

SUMO/MILO: = Animal

SENTIWN: P:0.0; N:0.5; O:0.5 |

Last Edit: Catalin Mihaila

--> [hyponym] *[n] organism:x, ființă:1.1

--> [holo_member] +[n] regnul_animal:x

<<< [hyponym] *[n] femelă:1, parte_femeiască:1

<<< [hyponym] *[n] mascul:1, parte_bărbătească:1

<<< [category_member] *[n] microorganism:1

<<< [hyponym] *[n] animal_cu_notocord:1

<<< [hyponym] *[n] nevertebrat:1

<<< [hyponym] *[n] larvă:1

<<< [category_member] *[n] corp:1, trup:1

<<< [mero_part] *[n] cap:1

<<< [hyponym] *[n] animal_domestic:1

<<< [hyponym] *[n] animal_de_muncă:1

<<< [hyponym] *[n] animal_de_pradă:1

<<< [hyponym] *[n] vânat:2.1

<<< [hyponym] *[n] animal_ierbivor:1

<<< [mero_part] *[n] cap:1.1.1, față:1

<<< [hyponym] *[n] vânat:2

<<< [category_member] [n] jertfă:4, jertfire:1, sacrificare:1, sacrificiu:4.1

<<< [hyponym] +[n] animal_de_companie:1

<<< [category_member] +[n] omenie:1.3

<<< [mero_portion] +[n] țesut_animal:1

<<< [hyponym] +[n] animal_de_cursă:1

<<< [hyponym] [n] pereche:4.1.1

<<< [category_member] [v] domestici:1.1, îmblânzi:1.1

<<< [category_member] *[a] adult:1.1, dezvoltat:1.1, matur:1.1

<<< [category_member] +[n] capcană:1.1, cursă:2.1.1, prinzătoare:1

<<< [category_member] [v] paște:1.3.2, pășuna:2.2, păstori:1.2

<<< [hyponym] *[n] prăsilă:2, progenitură:1.2, pui:1.4.x

<<< [hyponym] [n] pursânge:x

<<< [hyponym] [n] embrion:1, germen:1.2, ou:1.1, zigot:2

<<< [hyponym] [n] parazit:1

<<< [hyponym] [n] animal_marin:x, creatură_marină:x

<<< [category_member] [n] purtător:3.3

<<< [category_member] [n] parte:9

<<< [category_member] [n] urmărire:1.x

<<< [hyponym] [n] dobitoc:1

*[n] animal:1, animate being:1, beast:1, brute:2, creature:1, fauna:2

View Tree RevTree BCS1.2 BCS3 XML

POS: n ID: ENG20-00012748-n BCS: 1

Synonyms: animal:1, animate being:1, beast:1, brute:2, creature:1, fauna:2

Definition: a living organism characterized by voluntary movement

Domain: animals

Domain: biology

SUMO/MILO: = Animal

SENTIWN: P:0.0; N:0.5; O:0.5

--> [hyponym] *[n] organism:1, being:2

--> [holo_member] +[n] Animalia:1, kingdom Animalia:1, animal kingdom:1

--> [eng_derivative] *[v] make:3, create:1

--> [eng_derivative] [v] animalize:1, animalise:1

<<< [category_member] [a] crested:2, topknotted:1, tufted:3

<<< [category_member] [a] bone-covered:1

<<< [category_member] [a] free-swimming:1, unattached:3

<<< [category_member] [v] bone:2, debone:1

<<< [category_member] [n] sacrifice:4, ritual killing:1

<<< [category_member] [a] hispid:1

<<< [category_member] [v] domesticate:3, tame:5

<<< [category_member] [v] domesticate:2, domesticize:1, domesticise:1, reclaim:5, tame:4

<<< [category_member] [n] trailing:1, tracking:1

<<< [category_member] [a] flesh-eating:1, meat-eating:1, zoophagous:1

<<< [category_member] [a] plant-eating:1, phytophagic:1, phytophagous:1, phytophilous:1

<<< [category_member] [a] all-devouring:1

<<< [category_member] [a] insectivorous:1

<<< [category_member] [v] drench:2

<<< [hyponym] [n] critter:1

<<< [hyponym] [n] creepy-crawly:1

<<< [hyponym] [n] darter:2

<<< [hyponym] [n] peeper:3

<<< [hyponym] [n] poikilotherm:1, ectotherm:1

<<< [hyponym] [n] range animal:1

<<< [hyponym] [n] vermin:2

<<< [hyponym] [n] varmint:2, varment:1

<<< [hyponym] [n] scavenger:3

<<< [hyponym] *[n] work animal:1

<<< [hyponym] *[n] domestic animal:1

<<< [hyponym] [n] migrator:2

<<< [hyponym] [n] molter:1, moult:1

<<< [hyponym] +[n] pet:1

<<< [hyponym] [n] stayer:1

<<< [hyponym] [n] stunt:2

<<< [hyponym] [n] marine animal:1, marine creature:1, sea animal:1, sea creature:1

<<< [hyponym] *[n] female:1

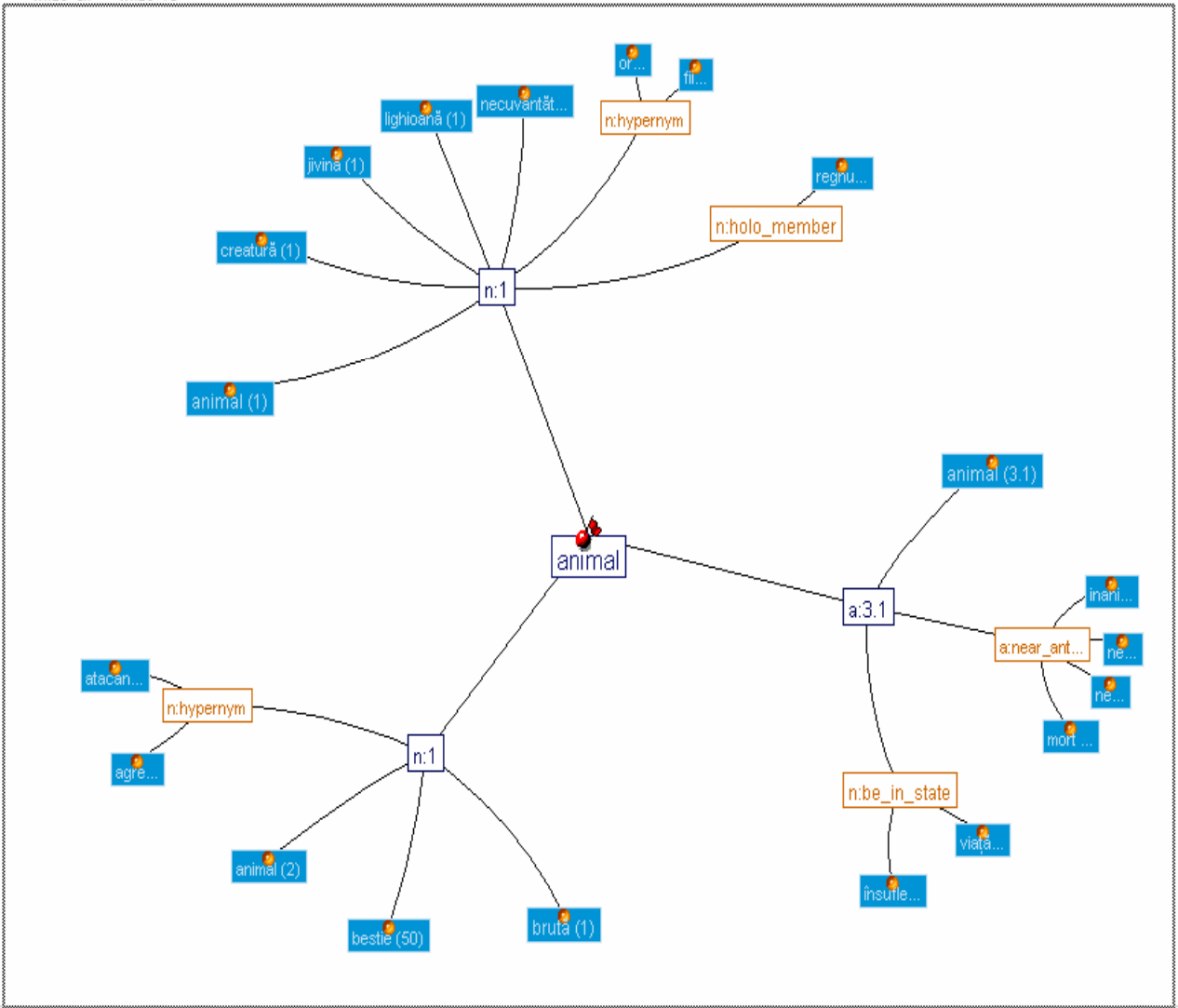
<<< [hyponym] *[n] male:1

Wordnet Browser

Help BalkaNet

animal Search

wn20-en wn20-ro



ENG20-00012748-n

- animal (1)
- creatură (1)
- jivină (1)
- lighioană (1)
- necuvântător (1)

DOMAIN: animals; SUMO: Animal=; STAMP: Catalin Mihaila

Ființă organizată, uni-sau pluricelulară, înzestrată cu facultatea de a simți și de a se mișca

ENG20-00124508-a

- animal (3.1)

DOMAIN: anatomy; SUMO: Animal=; STAMP: verginica

De animal, propriu animalelor.

ENG20-09218213-n

- brută (1)
- bestie (50)
- animal (2)

DOMAIN: law; SUMO: Human+; STAMP:

Om brutal, grosolan, josnic, care se poartă ca un animal

Preprocessing Parallel Corpora

- Cleaning up the documents in a consistent way for all languages, XCES level 1 independent annotation for each language
- Sentence alignment for all possible pairs
- Tokenization, tagging, lemmatization and chunking for all languages of interest (or for which resources are available)=>XCES ANA level 3
- Word Alignment for all language pairs or using a hub language =>XCES-Align level 3

XCES-ANA Tools

- Document cleaner (MS, TEXT, HTML ->XCES 1)
- SVM-based Sentence Aligner –language independent (~98% accuracy); with translation dictionaries, sentence alignment accuracy goes up next to 100%.
- Multilingual Tokenizer with specific language resources for 15 languages
- Trainable Tiered Tagging and Combined Classifiers (HMM+ME) and LM for the six languages of the Multext-EAST project
- Lemmatizers for the six languages of the Multext-EAST project
- XCES-ANA generator

XCES-Align (level 3) En-Ro-Bg

```
C:\Documents and Settings\Dan\Desktop\nou-jrc42002D0595.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Documents and Settings\Dan\Desktop\nou-jrc42002D0595.xml

<?xml version="1.0" ?>
<!DOCTYPE text (View Source for full doctype...)>
- <text id="jrc42002D0595">
- <body>
- <tu id="1">
- <seg lang="en">
- <s id="jrc42002D0595-en.1.en">
  <w lemma="decision" ana="Ncns">Decision</w>
  <w lemma="of" ana="Sp">of</w>
  <w lemma="the" ana="Dd">the</w>
  <w lemma="representative" ana="Ncnp">Representatives</w>
  <w lemma="of" ana="Sp">of</w>
  <w lemma="the" ana="Dd">the</w>
  <w lemma="government" ana="Ncnp">Governments</w>
  <w lemma="of" ana="Sp">of</w>
  <w lemma="the" ana="Dd">the</w>
  <w lemma="member" ana="Ncns">Member</w>
  <w lemma="state" ana="Ncnp">States</w>
  <c>,</c>
  <w lemma="meet" ana="Vmpp">meeting</w>
  <w lemma="within" ana="Sp">within</w>
  <w lemma="the" ana="Dd">the</w>
  <w lemma="council" ana="Ncns">Council</w>
  </s>
</seg>
- <seg lang="ro">
- <s id="jrc42002D0595-ro.1.ro">
  <w lemma="decizie" ana="Ncfsry">DECIZIA</w>
  <w lemma="reprezentant" ana="Ncmpoy">REPREZENTANȚILOR</w>
  <w lemma="guvern" ana="Ncfpoy">GUVERNELOR</w>
  <w lemma="stat" ana="Ncfpoy">STATELOR</w>
  <w lemma="membru" ana="Afpfp-n">MEMBRE</w>
  <c>,</c>
  <w lemma="reuni" ana="Vmp--pm">REUNIȚI</w>
  <w lemma="în" ana="Spsa">ÎN</w>
  <w lemma="consiliu" ana="Ncms-n">CONSILIU</w>
  </s>
</seg>
- <seg lang="bg">
- <s id="jrc42002D0595-bg.1.bg">
  <w lemma="решение" ana="Ncns-n">РЕШЕНИЕ</w>
  <w lemma="на" ana="Sp">НА</w>
  <w lemma="представител" ana="Ncmpr-y">ПРЕДСТАВИТЕЛИТЕ</w>
  <w lemma="на" ana="Sp">НА</w>
  <w lemma="правителство" ana="Ncnp-y">ПРАВИТЕЛСТВАТА</w>
  </s>
</seg>
</tu>
</body>
</text>
```

Lexical alignment of parallel documents

- Method: Reified lexical alignment (very high accuracy)
- Hub language – English: From En-X and En-Y alignments we automatically derive the X-Y alignment
- The derived alignments are used to generate translation models for the X-Y pair of languages and by bootstrapping (if necessary and linguistic expertise is available) to improve the initial X-Y alignment
- ***Reified Alignments - COWAL***
 - A bitext alignment is a set of lexical token pairs (*links*), each of them being characterized by a weighted feature structure the score of which should be higher than the acceptability threshold.

$$\left[\begin{array}{l} \alpha \text{ feat}_1: \text{val}_1 \\ \beta \text{ feat}_2: \text{val}_2 \\ \dots \\ \zeta \text{ feat}_n: \text{val}_n \end{array} \right]$$

$$\text{LinkScore} = \sum_{i=1}^n \text{CoefFeat}_i * \text{ScoreFeat}_i$$

Features characterizing a link

- A link <Token1 Token2> is characterized by a set of features, the values of which are real numbers in the [0,1] interval.
- **context independent features – CIF**, they refer to the tokens of the current link
 - cognate, translation equivalents (TE), POS-affinity, “obliqueness”, TE entropy
- **context dependent features – CDF**, they refer to the properties of the current link with respect to the rest of links in a bi-text.
 - strong and/or weak locality, number of links crossed, collocations
- Based on the values of a link’s features we compute for each possible link a global reliability score which is used to license or not a link in the final result.

ACL2005 Word Alignment Shared Task (En-Ro)

1. COWAL (F=73.90%, AER=26.10%)

The current version (July 2007) evaluated against our corrected ACL2005 GS:

COWAL (F=85.22%, AER=14.78%)

- ***Translation equivalents (TE)***

- YAWA uses an external bilingual lexicon (TREQ+RO&EN wordnets)
- MEBA uses GIZA++ generated candidates filtered with a log-likelihood threshold (11).
- For a pair of languages translation equivalents are computed in both directions. The value of the TE feature of a candidate link <TOKEN1 TOKEN2> is

$$1/2 (P_{TR}(TOKEN1, TOKEN2) + P_{TR}(TOKEN2, TOKEN1)).$$

- ***Translation Entropy Score (ES)***

- The entropy of a word's translation equivalents distribution proved to be an important hint on identifying highly reliable links (anchoring links)
- Skewed distributions are favored against uniform ones

$$ES(W) = 1 + \frac{\sum_{i=1}^N p(W, TR_i) * \log p(W, TR_i)}{\log N}$$

- For a link <A B>, the link feature value is 0.5(ES(A)+ES(B))

- ***Cognates (COGN)***

$$T_S = \alpha_1 \alpha_2 \dots \alpha_k \quad ; \quad T_T = \beta_1 \beta_2 \dots \beta_m$$

if α_i and β_j are the matching characters, &

if $\delta(\alpha_i)$ is the distance (in chars of T_S) from the previous matching α , &

if $\delta(\beta_j)$ is the distance (in chars of T_T) from the previous matching β

then

$$SYM(T_S, T_T) = \begin{cases} \frac{\sum_{i=1}^q \frac{2}{1 + |\delta(\alpha_i) - \delta(\beta_i)|}}{k + m} & \text{if } q > 2 \\ 0 & \text{if } q \leq 2 \end{cases}$$

$$COGN(T_S, T_T) = \begin{cases} 1 & \text{if } SYM(T_S, T_T) > \text{Threshold} \\ 0 & \text{otherwise} \end{cases}$$

- ***Part-of-speech affinity (PA)***

The translated words tend to keep their part-of-speech and when they have different POSes, this is not arbitrary. The information was computed based on a gold standard (GS2003), in both directions (source-target and target-source).

$$\text{For a link } \langle A, B \rangle \quad PA = 0.5 * (P(\text{cat}(A) | \text{cat}(B)) + P(\text{cat}(B) | \text{cat}(A)))$$

- ***Collocation***

- Bi-gram lists (only content words) were built from each monolingual part of the training corpus, using the log-likelihood score (threshold of 10) and minimal occurrence frequency (3) for candidates filtering. Collocation probabilities are estimated for each surviving bi-gram.
- If neither token of a candidate link has a relevant collocation score with the tokens in its neighborhood, the link value of this feature is 0. Otherwise the value is 1. Competing links (starting or finishing in the same token) for YAWA are licensed only and only if at least one of them have a non-null collocation score.

- ***Obliqueness***

- Each token in both sides of a bi-text is characterized by a position index, computed as the ratio between the relative position in the sentence and the length of the sentence. The absolute value of the difference between tokens' position indexes subtracted from 1, gives the link's "*obliqueness*" $OBL(<SW_i, TW_j>)$.

$$OBL(< SW_i, TW_j >) = 1 - \left| \frac{i}{length(Sent_S)} - \frac{j}{length(Sent_T)} \right|$$

- ***Locality***

- When the dependency chunking module is available, and the chunks are aligned via the linking of their constituents, the new candidate links starting in one chunk should finish in the aligned chunk (**strong locality**).
 - ***Strong Locality:EM and CLAM Combined Linkers***
 - We have modified the EM algorithm of IBM-1 to work on a ‘bitext’ that contains the source sentence and a replica of it as the target:
 - » disregard the NULL alignment links
 - » disregard words that are on the same position
 - LAM introduced by Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction*. PhD thesis, Dept of Computer Science and Electrical Engineering, MIT (subject to planarity restriction)
 - Constrained LAM: a link is rejected if it does not pass any of the linking rules of a language: for instance the number agreement
- When the dependency chunking is not available, the locality is judged in a variable length window depending on the length of the current aligned sentences (**weak locality**)

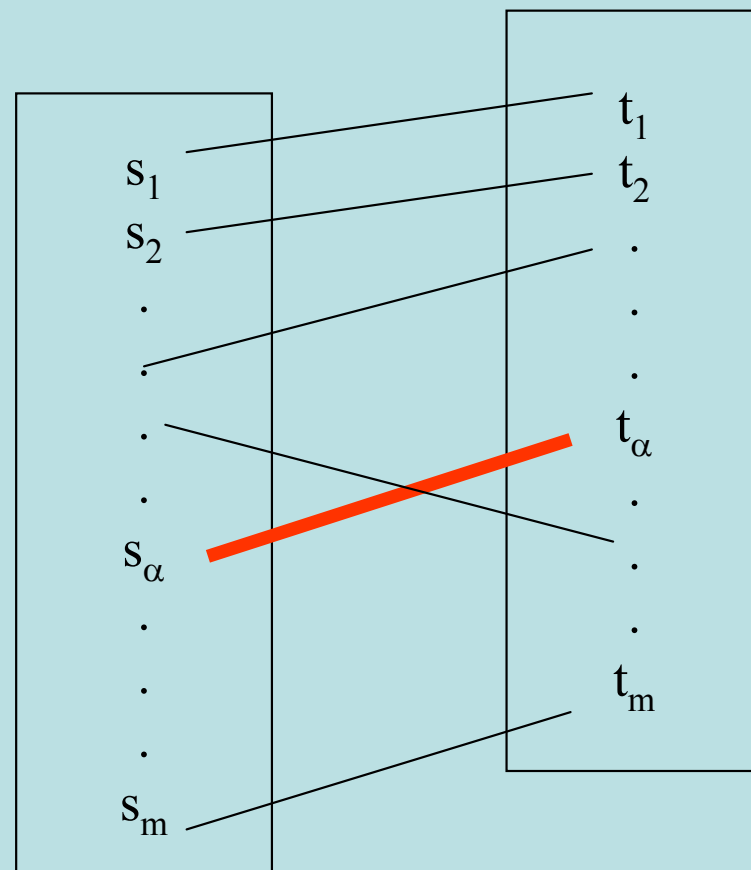
- **Weak Locality**

When chunking/dependency links information is not available, the link localization is judged against a window containing m links. The value of m depends on the aligned sentences length. The window is centered on the candidate link.

$$LOC = \frac{1}{m} \sum_{k=1}^m \frac{\min(|s_\alpha - s_k|, |t_\alpha - t_k|)}{\max(|s_\alpha - s_k|, |t_\alpha - t_k|)}$$

- **Combining classifiers**

If multiple classifiers are comparable, and if they do m not make similar errors, combining their classifications is always better than the Individual classifications.



COWAL

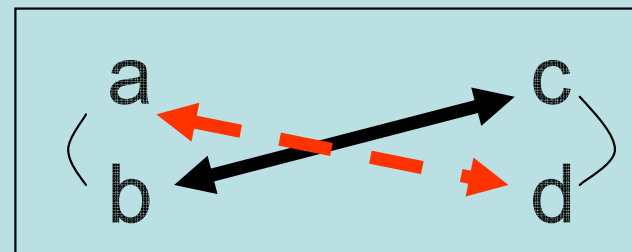
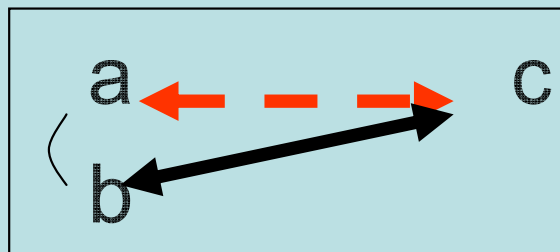
- An integrated platform that takes two parallel raw texts and produces their alignment
 - **basic modules**: collocations detector, tokenizers, lemmatizers, POS-taggers, two or more comparable word-aligners (YAWA, MEBA), GIZA++ translation model builder, alignment combiner,
 - **optional modules** : sentence aligner,, dependency “linkers, chunkers and bilingual dictionaries (Ro-En aligned wordnets)
 - The platform also includes an XML generator (XCES schema compliant), an alignment viewer & editor, and a WSD based on WA and aligned wordnets.

Combining the Alignments

- COWAL filters the reunion of the alignments. The filtering is achieved by a SVM classifier (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) trained on our version of the GS2005 (for positive examples) and the differences among the basic alignments (YAWA, MEBA) and the GS2005 (for negative examples);
- The SVM classifier (LIBSVM (Fan et al., 2005) uses the default parameters: C-SVC classification (soft margin classifier) and RBF kernel (Radial Basis Function $K(x, y) = e^{-\gamma \|x-y\|^2}$)
- Features used for the training (10-fold validation; about 7000 good examples and 7000 bad examples) :
TE(S,T), TE(T,S), OBL(S,T), LOC(S,T), PA(S,T), PA(T,S)
The links labeled as *incorrect* links were removed from the merged alignments.

Heuristics for improving the alignment (1)

- The words unaligned in the previous step may get links via their aligned dependents (HLP: *Head Linking Projection heuristics*): if b is aligned to c and b is linked to a , link a to c , unless there exist d in the same chunk with c , linked or not to it, and the POS category of d has a significant affinity with the category of a .

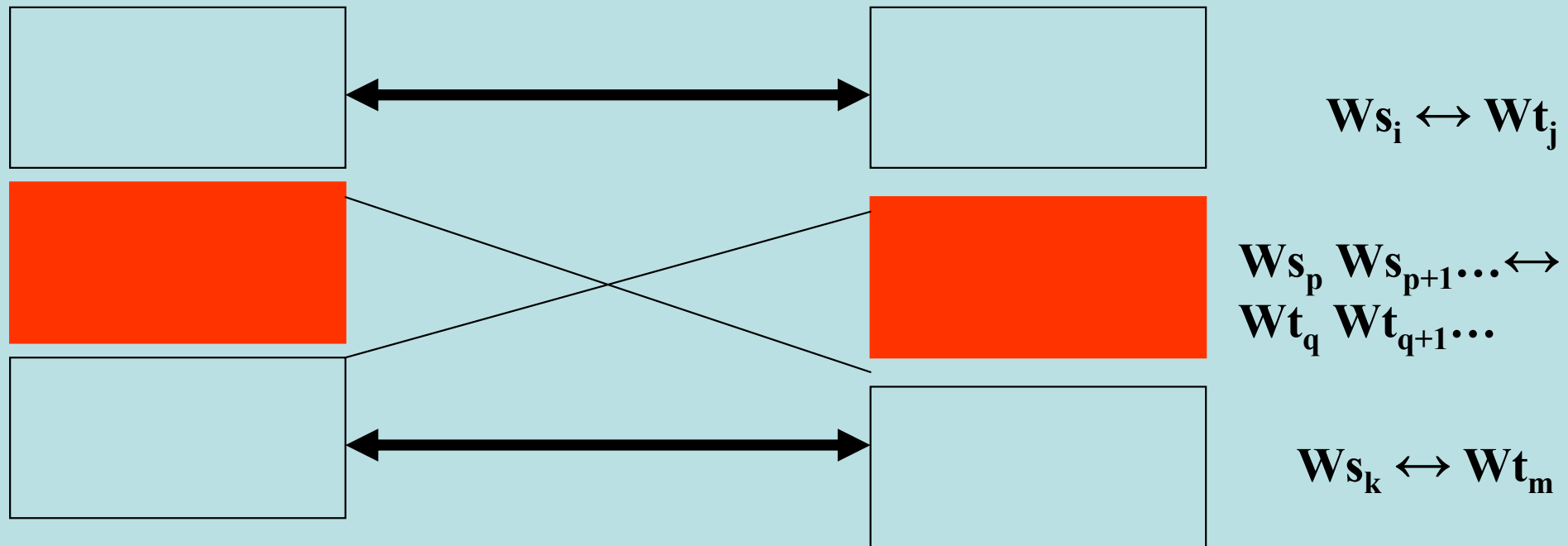


- Alignment of sequences of words surrounded by the aligned chunks
- Filtering out improbable links (e.g.links that cross many other links)

Heuristics for improving the alignment (2)

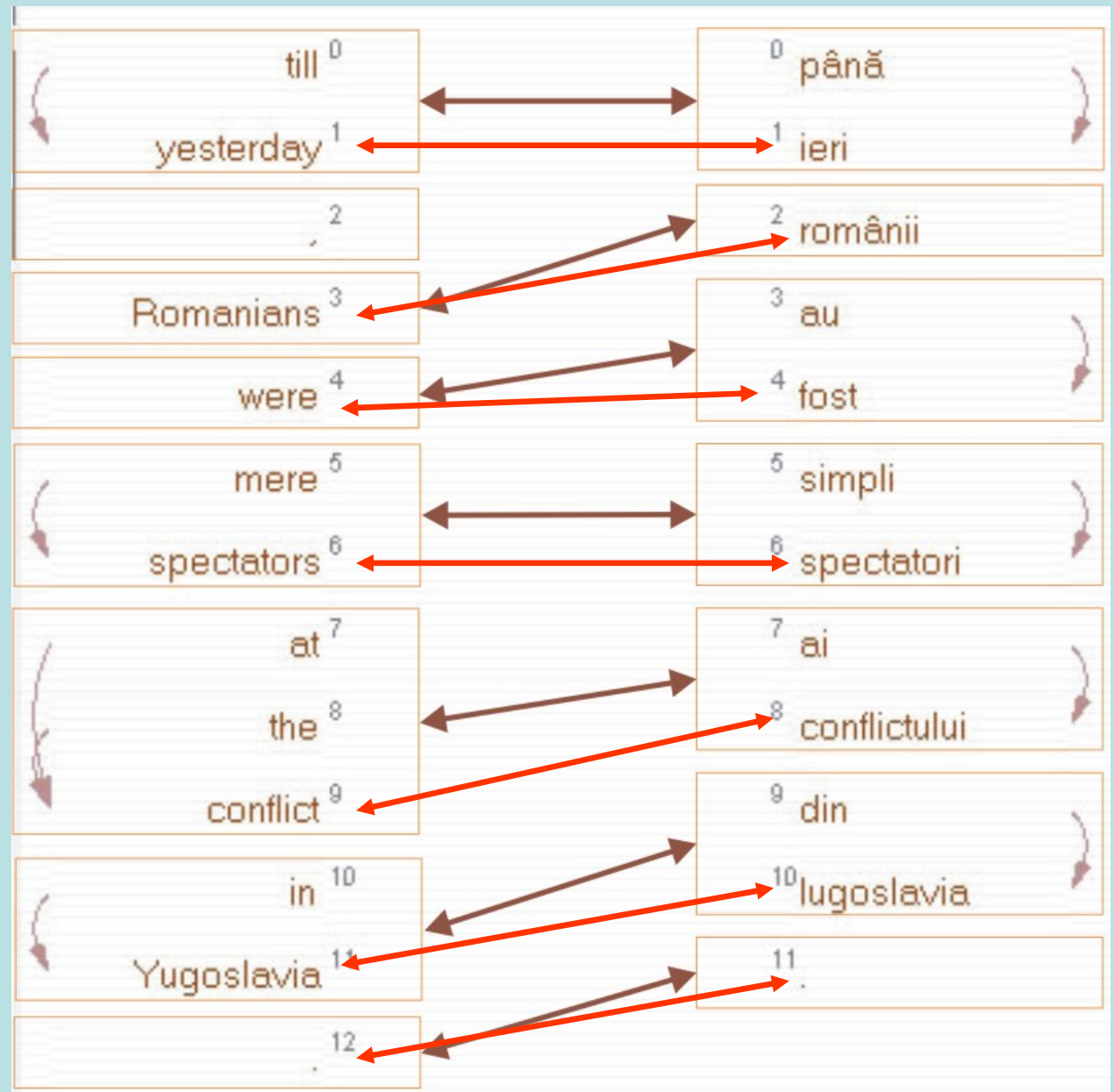
- Unaligned chunks surrounded by aligned chunks get probable phrase alignment:

SL \longleftrightarrow TL



Dependency chunks & Translation Model

- Regular expressions defined over the POS tags and dependency links
- Non-recursive chunks
- Chunk alignment based on their aligned constituents (one or more).



En-Ro Word Alignment

MTKit

Aligner Save Align Viewer

Aligner nou-jrc42002D0595.align | nou-jrc42002D0595bg.align | nou-jrc42002D0595robg.align

16

nou-jrc4...	
01-01	This 1 ————— 1 Prezenta
02-02	
03-03	Decision 2 ————— 2 decizie
03-04	
03-05	shall 3 ————— 3 intră
04-03	
04-04	enter 4 ————— 4 în
04-05	
05-03	into 5 ————— 5 vigoare
05-04	
05-05	force 6 ————— 6 în
06-03	
06-04	on 7 ————— 7 ziua
06-05	
07-06	the 8 ————— 8 următoare
08-07	
09-07	day 9 ————— 9 publicării
10-08	
11-10	following 10 ————— 10 sale
12-09	
13-11	its 11 ————— 11 în
14-12	
15-13	publication 12 ————— 12 Jurnalul
16-12	
17-14	in 13 ————— 13 Oficial
18-15	
19-16	the 14 ————— 14 al
20-15	
21-17	Official 15 ————— 15 Comunităților
	Journal 16 ————— 16 Europene
	of 17 ————— 17
	the 18 —————
	European 19 —————
	Communities 20 —————
	21 —————

Ready.

Properties

Main

Ana	Ncns
Chunk	
Lemma	journal
Ortho	Journal
TAna	1

WordNet

ILI	ENG20-06189977-n
SUMO	Periodical
WNS	2

Ana

WordNet

Word: Journal (journal)

Language: en

- n. 1. diary(1), journal(1)
- n. 2. journal(2)
- n. 3. daybook(1), journal(3)
- n. 4. journal(4)

Morpheme

- Phrase
 - Text
 - MotionPicture
 - FactualText
 - FictionalText
 - Book
 - Summary
 - Series
 - Periodical
 - Article

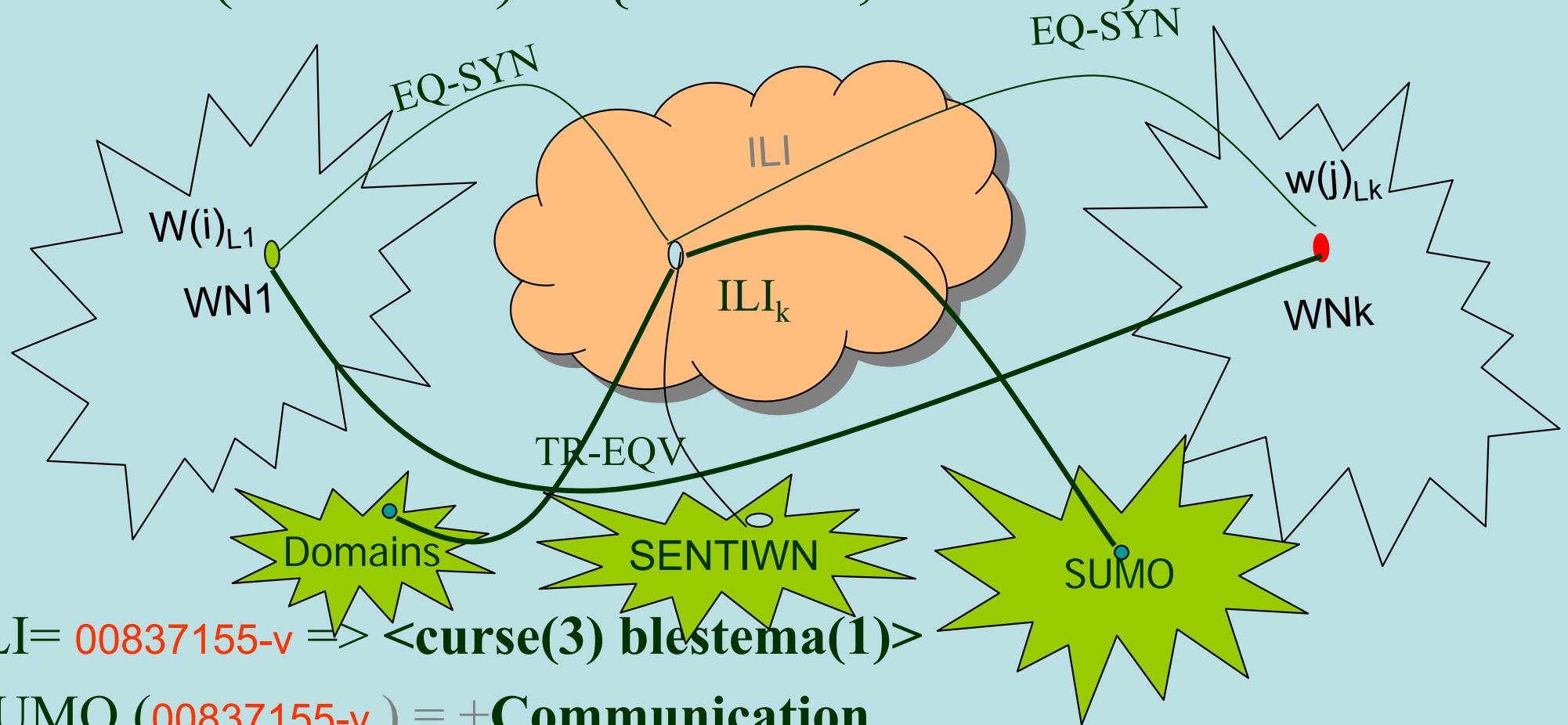
[ENG20-06189977-n] a periodical dedicated to a particular subject

SUMO documentation: A &%Series whose elements are published separately and on a periodic basis.

Example (I): <lamp lampă>

PWN2.0 (curse) = {00836766-v, 00837493-v, 00837155-v, 00997108-v}

RoWN (blestema) = {00837155-v, 00837155-v}



ILI= 00837155-v => <curse(3) blestema(1)>

SUMO (00837155-v) = +Communication

SENTIWN (00837155-v) = P:0.0; N:0.625; O:0.375

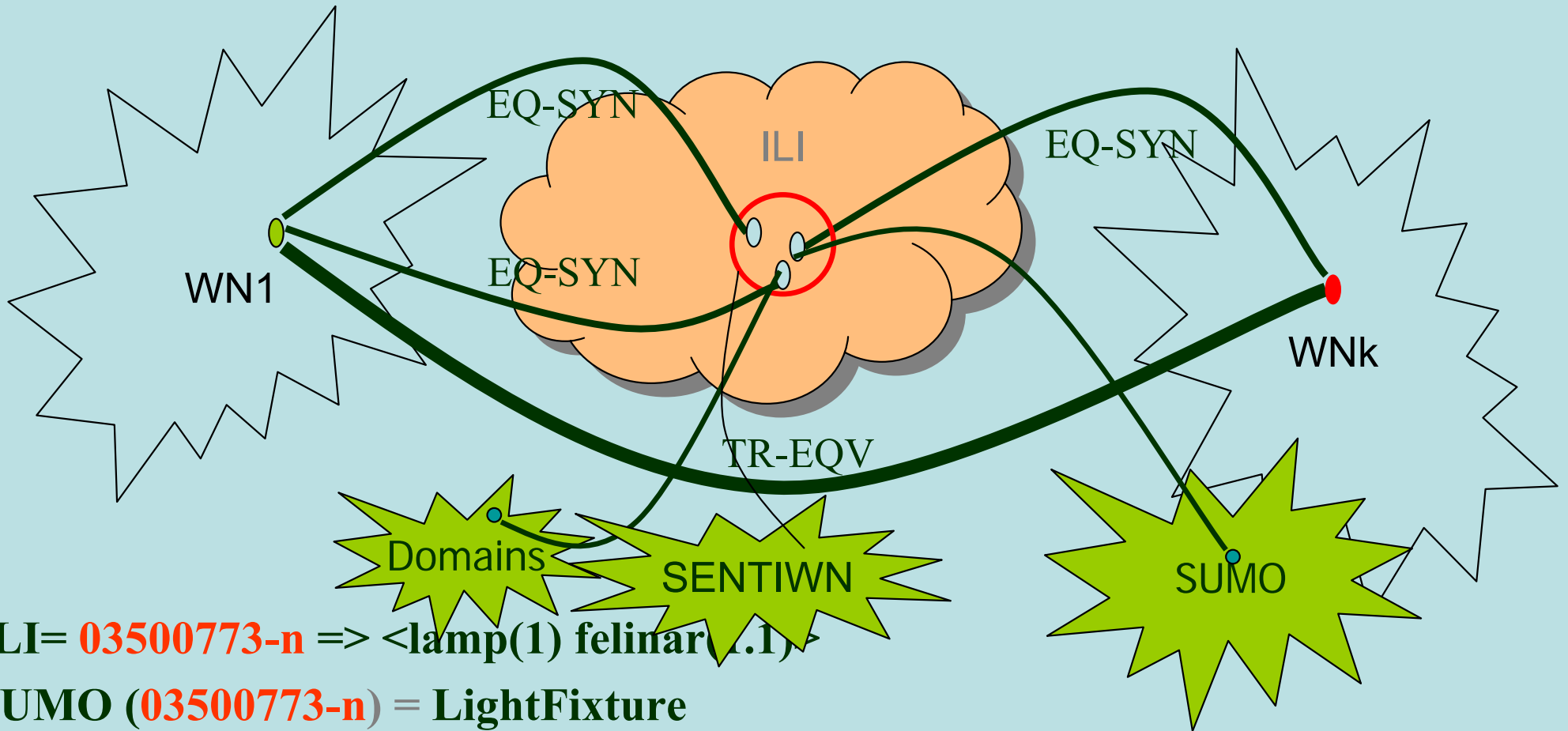
DOMAINS (00837155-v) = factotum

Example (II): <lamp felinar>

PWN2.0 (lamp) = {03500372-n, 03500773-n}

RoWN (felinar) = {003505057-n}

δ (03500372-n, 003505057-n)=0.5 δ (03500373-n, 003505057-n)=0.125



ILI= 03500773-n => <lamp(1) felinar(1.1)>

SUMO (03500773-n) = LightFixture

SUMO (003505057-n) = +IlluminationDevice

SENTIWN (03500773-n, 003505057-n) = P:0.0;N:0.0;O:1

DOMAINS (03500773-n, 003505057-n) = factotum

Sentence Subjectivity Scorer

- A very naïve implementation:
 - For each sentence in each language add the P, N and O figures of each word

The stuff(1) was(1) like (3) nitric_acid(1) , and moreover(1) , in swallowing (1) it one had (1) the sensation (1) of being hit (4) on the back_of_the_head (1) with a rubber(1) club(3).

Sentence_1 score: P:0,031;N:0.042;O:0.927

OpinionFinder says:

autoclass1="subj" autoclass2="subj" diff="30.8"⁴⁴

What's wrong with this naïve scorer?

- It doesn't consider the valency shifters.

*The stuff(1) **was_like***

nitric_acid(1)... P:0;N:0:O:1=>P:0.5;N:0.5;O:0

...

***had_sensation** of being hit (4) on the
back_of_the_head (1) with a rubber(1) club(3).*

With valency shifters considered, either the SO or the PN or both polarities are switched.

Sentence_1 score: P:0,063;N:0.563;O:0.375

Now this is in line with OpinionFinder!

Sentence Subjectivity Scorer

- A very naïve implementation:
 - For each sentence in each language add the P, N and O figures of each word

He has(1) no(1) merits(1).

P:0.0;N:0.0;O:1

P:0.25;N:0.25;O:0.5

P:0.625;N:0.0;O:0.375

Sentence_1 score: P:0.292;N:0.083;O:0.625

He has(1) all the merits(1).

P:0.0;N:0.0;O:1

P:0.0;N:0.0;O:1

P:0.625;N:0.0;O:0.375

Sentence_2 score: P:0.208;N:0.0;O:0.792

Sentence Subjectivity Scorer

- A very naïve implementation:
 - For each sentence in each language add the P, N and O figures of each word

He has(1) no(1) merits(1).

P:0.0;N:0.0;O:1

P:0.25;N:0.25;O:0.5

P:0.625;N:0.0;O:0.375

Sentence_1 score: P:0.292;N:0.083;O:0.625

He has(1) all the merits(1).

P:0.0;N:0.0;O:1

P:0.0;N:0.0;O:1

P:0.625;N:0.0;O:0.375

Sentence_2 score: P:0.208;N:0.0;O:0.792

Still naïve implementation:

For each sentence in each language add the P, N and O figures of each word unless it was preceded by a valency-shifter word (most of them were extracted from Wordnet based on a hand-made seed list).

He has(1) no(1) merits(1).

P:0.0;N:0.0;O:1

P:0.25;N:0.25;O:0.5

P:0.625;N:0.0;O:0.375 ... (cancelled by the “no” VS)

Sentence_1' score: P:0.125;N:0.125;O:0.75

He has(1) all the merits(1).

P:0.0;N:0.0;O:1

P:0.0;N:0.0;O:1

P:0.625;N:0.0;O:0.375

Sentence_2' score: P:0.208;N:0.0;O:0.792

He has(1) no(1) merits(1).

Sentence_1 score: P:0.292;N:0.083;O:0.625

Sentence_1' score: P:0.125;N:0.125;O:0.75

He has(1) all the merits(1).

Sentence_2 score: P:0.208;N:0.0;O:0.792

Sentence_2' score: P:0.208;N:0.0;O:0.792

A More Realistic Scorer and Classifier

Do you really want to get everything for free?

This is your job!

If your classifier does better than other OM (e.g. OpinionFinder), please let me know.

Otherwise, contact me in two months from now on and I'll tell you the rest of the story...