TLN-LAB

Automatic summarisation with NASARI

Daniele Radicioni

BabelNet

di.unito.it
DIPARTIMENTO
DI INFORMATICA

credits

the following slides have been built on materials from:

Roberto Navigli and Simone Paolo Ponzetto (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216-225). Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, Artificial Intelligence, 193 (2012) 217-250.

Number of monosemous and polysemous words by part of speech (verbs, adjectives and adverbs are the same as in WordNet 3.0).

POS	Monosemous words	Polysemous words
Noun	22,763,265	1,134,857
Verb	6,277	5,252
Adjective	1,503	4,976
Adverb	3,748	733
Total	22,789,793	1,145,818

Composition of Babel synsets: number of synonyms from the English WordNet, Wikipedia pages and translations, as well as translations of WordNet's monosemous words and SemCor's sense annotations.

		English	Catalan	French	German	Italian	Spanish	Total
English WordNet		206,978	–	–	–	–	–	206,978
Wikipedia	pages	2,955,552	123,101	524,897	506,892	404,153	349,375	4,863,970
	redirections	3,388,049	105,147	617,379	456,977	217,963	404,009	5,189,524
	translations	–	3,445,273	2,844,645	2,841,914	3,046,323	3,083,365	15,261,520
WordNet	monosemous	–	97,327	97,680	97,852	98,089	97,435	488,383
	SemCor	–	6,852	6,855	6,850	6,856	6,855	34,268
Total		6,550,579	3,777,700	4,091,456	3,910,485	3,773,384	3,941,039	26,044,643

Multilingual lexical resources

- Lexical knowledge is available in many different forms, ranging from unstructured terminologies (i.e., lists of terms), to glossaries (e.g., Web-derived domain glossaries), thesauri, machine-readable dictionaries and computational lexicons and ontologies, such as WordNet and Cyc.
  - However, building such resources manually is an onerous task.
  - It requires dozens of years, and has to be repeated from scratch for each new language.

Multilingual lexical resources

- Further, it has to be added the cost of interlinking the resources across languages and domains.
  - Manual efforts of this kind include EuroWordNet, MultiWordNet, BalkaNet, and others.
- However, resources for non-English languages often have much poorer coverage.
  - As a result, an obvious bias exists towards conducting research in resource-rich languages such as English.

Wikipedia

- Much work in the literature has been devoted to the extraction of structured information from Wikipedia, including extracting lexical and semantic relations between concepts, factual information, and transforming the Web encyclopaedia into a full-fledged semantic network.
  - One major feature of Wikipedia is its richness of explicit and implicit semantic knowledge, mostly about *named entities* (e.g., *Apple* as a company).

Wikipedia and WordNet

- However, Wikipedia's encyclopaedic nature is also a major limit, in that it lacks full coverage for the lexicographic senses of a given lemma
  - e.g., the *apple* 'fruit' and 'tree' senses are merged into one single meaning.
- Such a lexical coverage, instead, can be provided by a highly-structured computational lexicon such as WordNet.

BabelNet

- BabelNet aims at providing an “encyclopaedic dictionary” by merging WordNet and Wikipedia.

WordNet: notazione

- A concept in WordNet is represented as a synonym set (called *synset*).
  - For instance, the concept of *play* as a dramatic work is expressed by the following synset:

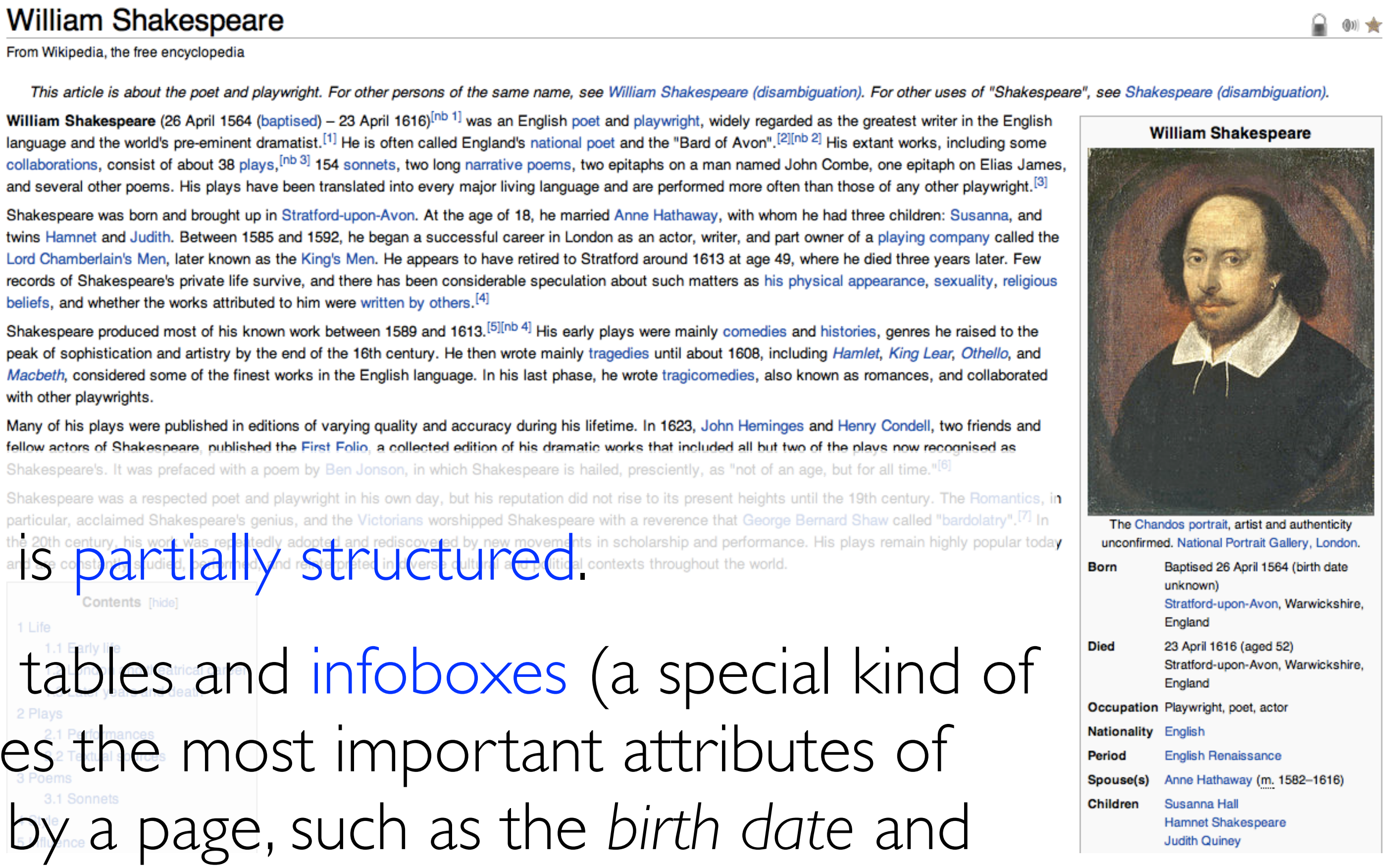
$$\{\text{play}_n^1, \text{drama}_n^1, \text{dramatic play}_n^1\}$$

- each word's subscript and superscript indicate its part of speech (e.g., *n* stands for noun) and sense number, respectively.
- Words can be polysemous and therefore the same word, e.g., *play*, can appear in more than one synset.

Wikipedia

- Wikipedia is a multilingual Web-based encyclopaedia.
  - It is a collaborative open source medium edited by volunteers to provide a very large wide-coverage repository of encyclopaedic knowledge.
- Each article in Wikipedia is represented as a page (Wikipage) and presents information about a specific *concept* (e.g., *Play (theatre)*) or named entity (e.g., *William Shakespeare*).
- The title of a Wikipage (e.g., *Play (theatre)*) contains an optional label in parentheses which specifies its meaning if the lemma is ambiguous (e.g., *theatre* vs. *activity*)

Wikipedia



William Shakespeare

From Wikipedia, the free encyclopedia

This article is about the poet and playwright. For other persons of the same name, see William Shakespeare (disambiguation). For other uses of "Shakespeare", see Shakespeare (disambiguation).

William Shakespeare (26 April 1564 (baptised) – 23 April 1616)[nb 1] was an English poet and playwright, widely regarded as the greatest writer in the English language and the world's pre-eminent dramatist.[1] He is often called England's national poet and the "Bard of Avon".[2][nb 2] His extant works, including some collaborations, consist of about 38 plays,[nb 3] 154 sonnets, two long narrative poems, two epitaphs on a man named John Combe, one epitaph on Elias James, and several other poems. His plays have been translated into every major living language and are performed more often than those of any other playwright.[3]

Shakespeare was born and brought up in Stratford-upon-Avon. At the age of 18, he married Anne Hathaway, with whom he had three children: Susanna, and twins Hamnet and Judith. Between 1585 and 1592, he began a successful career in London as an actor, writer, and part owner of a playing company called the Lord Chamberlain's Men, later known as the King's Men. He appears to have retired to Stratford around 1613 at age 49, where he died three years later. Few records of Shakespeare's private life survive, and there has been considerable speculation about such matters as his physical appearance, sexuality, religious beliefs, and whether the works attributed to him were written by others.[4]

Shakespeare produced most of his known work between 1589 and 1613.[5][nb 4] His early plays were mainly comedies and histories, genres he raised to the peak of sophistication and artistry by the end of the 16th century. He then wrote mainly tragedies until about 1608, including *Hamlet*, *King Lear*, *Othello*, and *Macbeth*, considered some of the finest works in the English language. In his last phase, he wrote tragicomedies, also known as romances, and collaborated with other playwrights.

Many of his plays were published in editions of varying quality and accuracy during his lifetime. In 1623, John Heminges and Henry Condell, two friends and fellow actors of Shakespeare, published the First Folio, a collected edition of his dramatic works that included all but two of the plays now recognised as Shakespeare's. It was prefaced with a poem by Ben Jonson, in which Shakespeare is hailed, presciently, as "not of an age, but for all time."[6]

Shakespeare was a respected poet and playwright in his own day, but his reputation did not rise to its present heights until the 19th century. The Romantics, in particular, acclaimed Shakespeare's genius, and the Victorians worshipped Shakespeare with a reverence that George Bernard Shaw called "bardolatry".[7] In the 20th century, his work was repeatedly adopted and rediscovered by new movements in scholarship and performance. His plays remain highly popular today and are constantly studied, performed, and reinterpreted in diverse cultural and political contexts throughout the world.

Contents [hide]

William Shakespeare

The Chandos portrait, artist and authenticity unconfirmed. National Portrait Gallery, London.

Born	Baptised 26 April 1564 (birth date unknown) Stratford-upon-Avon, Warwickshire, England
Died	23 April 1616 (aged 52) Stratford-upon-Avon, Warwickshire, England
Occupation	Playwright, poet, actor
Nationality	English
Period	English Renaissance
Spouse(s)	Anne Hathaway (m. 1582–1616)
Children	Susanna Hall Hamnet Shakespeare Judith Quiney

- The text in Wikipedia is partially structured.
- Some Wikipages have tables and infoboxes (a special kind of table which summarizes the most important attributes of the entity referred to by a page, such as the *birth date* and biographical details of a playwright like William Shakespeare).
- Additionally, various relations exist between the pages themselves.

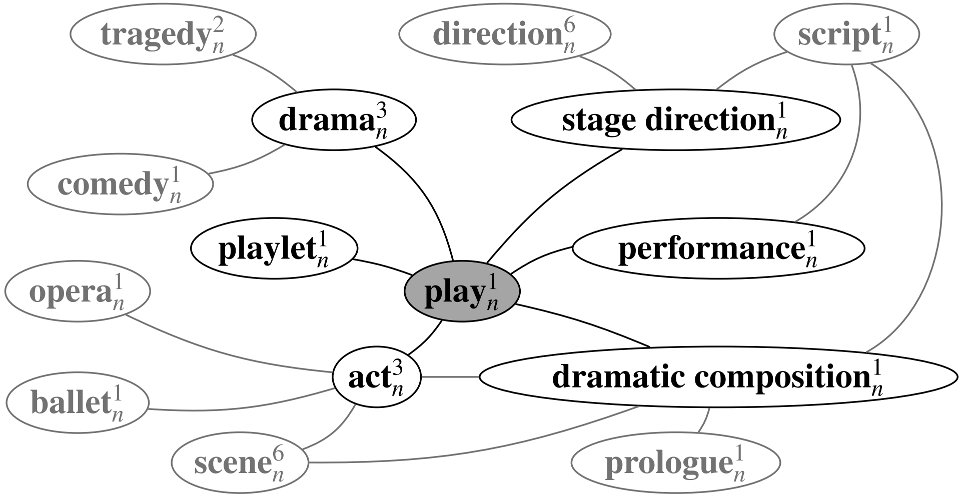
Relations between pages

- *Redirect pages*: used to forward to the Wikipage containing the actual information about a concept of interest.
  - This is used to point alternative expressions for a concept to the same entry, and thus models synonymy.
  - For instance, *Stageplay* and *Theatrical play* both redirect to *Play (theatre)*.
- *Disambiguation pages*: These pages collect links for possible concepts an arbitrary expression could be referred to;
  - e.g., *Play* links to both pages *Play (theatre)* and *Play (activity)*.

Relations between the pages

- *Internal links*: Wikipages typically contain hypertext linked to other Wikipages, which refer to related concepts.
  - For instance, *Play (theatre)* links to *Literature*, *Playwright*, *Dialogue*, etc., whereas *Play (activity)* points to *Socialization*, *Game*, *Recreation*, and so on.
- *Inter-language links*: Wikipages also provide links to the corresponding concepts contained within wikipedias in other languages;
  - e.g., the English Wikipage *Play (theatre)* links to the Italian *Dramma* and German *Bühnenwerk*.
- *Categories*: Wikipages can be assigned to one or more categories, i.e., special pages used to encode topics;
  - e.g., *Play (theatre)* is categorized under *THEATRE*, *DRAMA*, *LITERATURE*, etc.

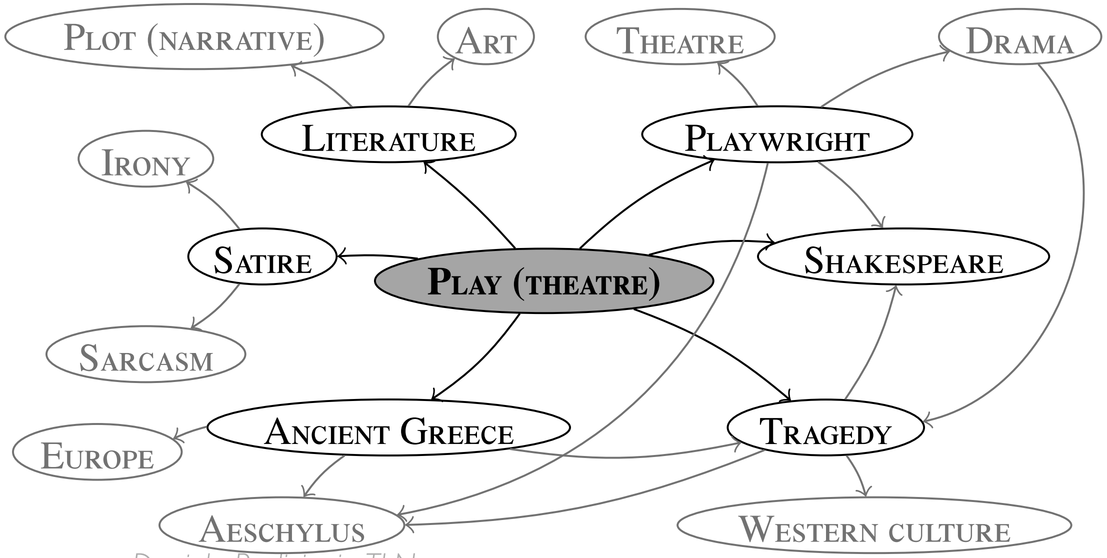
WordNet



Both can be viewed as graphs. In WordNet, nodes are synsets and edges lexical and semantic relations between synsets.

Wikipedia

in Wikipedia, nodes are Wikipages and edges the hyperlinks between them.



BabelNet and Babel synsets

- BabelNet encodes knowledge as a labeled directed graph $G = (V,E)$ where
  - V is the set of nodes (i.e., concepts such as *play* and named entities such as *Shakespeare*) and
  - $E \subseteq V \times R \times V$ is the set of edges connecting pairs of concepts (e.g., *play is-a dramatic composition*). Each edge is labeled with a semantic relation from R, i.e., {*is-a, part-of, ...,* ε}, where ε denotes an unspecified semantic relation.
- each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., {play_{en}, $\mathit{Theaterstück}_{de}$, dramma_{it}, obra_{es}, ... , $\mathit{pièce\ de\ théâtre}_{fr}$}.
  - We call such multilingually lexicalized concepts Babel synsets.

building the graph

- to build the BabelNet graph, the information collected is:
  - From WordNet, all available word senses (as *concepts*) and all the lexical and semantic pointers between synsets (as *relations*);
  - From Wikipedia, all encyclopedic entries (i.e., Wikipages, as *concepts*) and semantically unspecified *relations* from hyper-linked text.

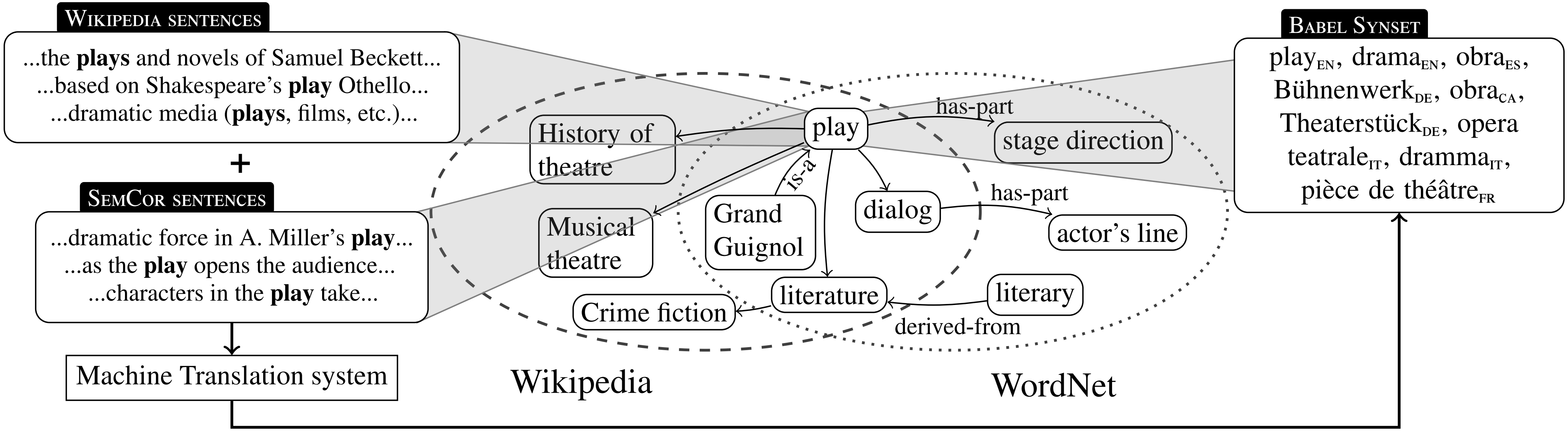
building the graph

- WordNet and Wikipedia can overlap both in terms of concepts and relations
  - in order to provide a unified resource, the intersection of these two knowledge sources is merged.
- Next, to enable multilinguality, the lexical realizations of the available concepts in different languages are collected.
- Finally, multilingual Babel synsets are connected by establishing semantic relations between them.

methodology

1. WordNet and Wikipedia are combined by automatically acquiring a mapping between WordNet senses and Wikipages.
2. Multilingual lexicalizations of the available concepts are harvested (i.e., Babel synsets) by using (*a*) the human-generated translations provided by Wikipedia (the so-called inter-language links), as well as (*b*) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.
3. The relations between Babel synsets are established by collecting all relations found in WordNet, as well as all wikipedias in the languages of interest: the strength of association between synsets is computed through a measure of relatedness based on the Dice coefficient.

methodology


WIKIPEDIA SENTENCES
...the **plays** and novels of Samuel Beckett...
...based on Shakespeare's **play** Othello...
...dramatic media (**plays**, films, etc.)...
+
SEMCOR SENTENCES
...dramatic force in A. Miller's **play**...
...as the **play** opens the audience...
...characters in the **play** take...
Machine Translation system
History of theatre
play
has-part
stage direction
is-a
Musical theatre
Grand Guignol
dialog
has-part
actor's line
literature
literary
derived-from
Crime fiction
Wikipedia
WordNet
BABEL SYNSET
play_EN, drama_EN, obra_ES, Bühnenwerk_DE, obra_CA, Theaterstück_DE, opera teatrale_IT, dramma_IT, pièce de théâtre_FR

unlabeled edges are obtained from links in the Wikipages (e.g., *Play (theatre)* links to *Musical theatre*), whereas labeled ones from WordNet (e.g., $play_n^1$ has-part stage $direction_n^1$).

NASARI

dı.unıto.ıt
DIPARTIMENTO
DI INFORMATICA

credits

- the following slides have been built based on

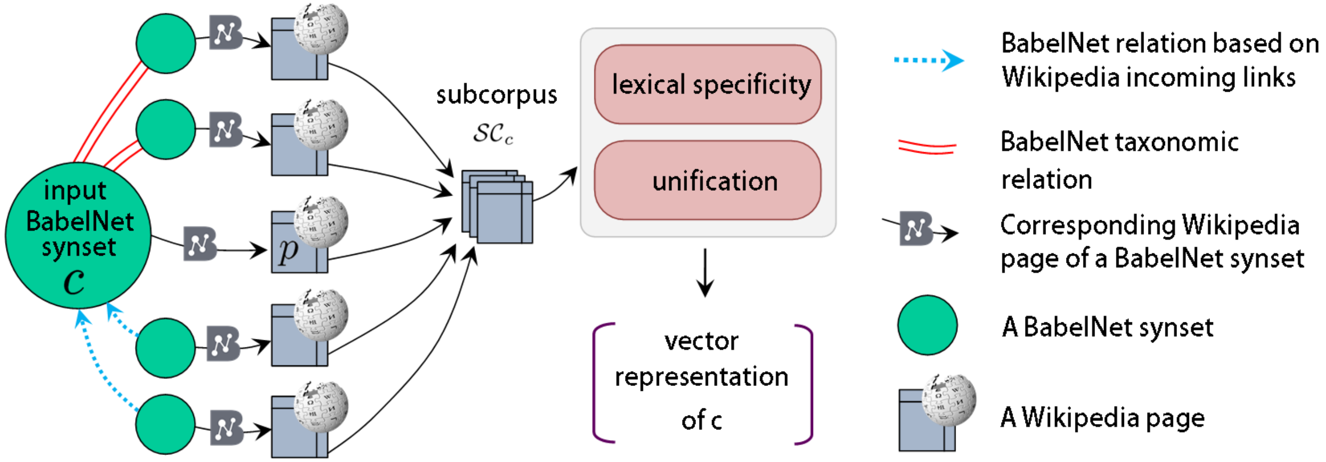
  José Camacho-Collados, Mohammad Taher Pilehvar and Roberto Navigli. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL 2015)*, Denver, USA, pp. 567-577, 2015

  José Camacho-Collados, Mohammad Taher Pilehvar and Roberto Navigli. A Unified Multilingual Semantic Representation of Concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Beijing, China, July 27-29, pp. 741-751, 2015

NASARI

- A Novel Approach to a Semantically-Aware Representation of Items
  - The prevailing methods for the computation of a vector space representation are based on distributional semantics.
  - These approaches are unable to model individual word senses or concepts, since they conflate different meanings of a word into a single vectorial representation.
    - Chen et al. (2014) addressed this issue and obtained vectors for individual word senses by leveraging WordNet glosses.

MUFFIN (Multilingual, UniFied and Flexible INterpretation)


input BabelNet synset
c
p
subcorpus
$\mathcal{SC}_c$
lexical specificity
unification
vector representation of c
BabelNet relation based on Wikipedia incoming links
BabelNet taxonomic relation
Corresponding Wikipedia page of a BabelNet synset
A BabelNet synset
A Wikipedia page

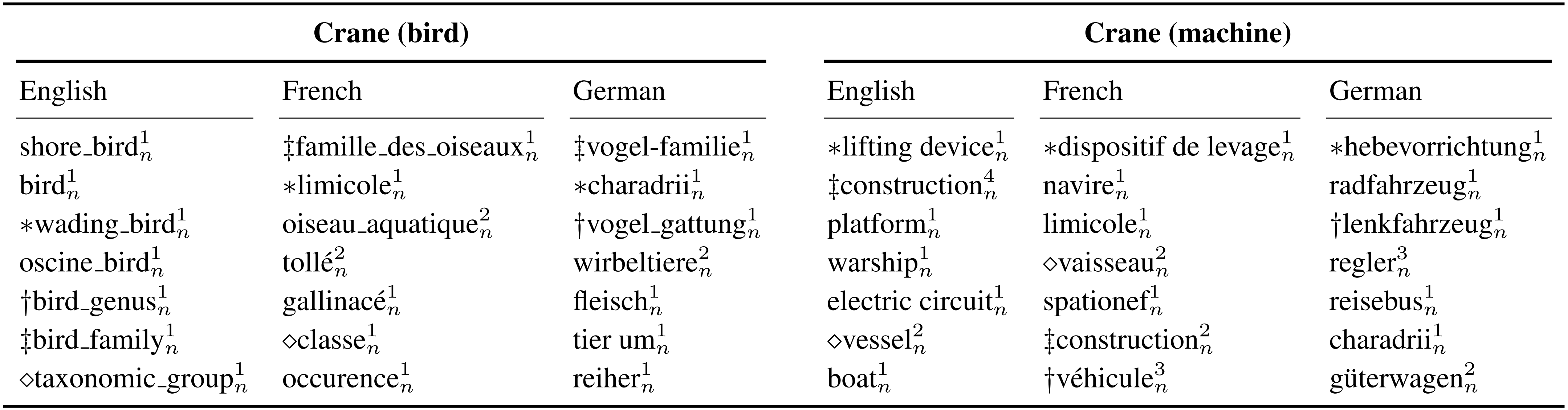
- For the given synset the contextual information is gathered from Wikipedia by exploiting knowledge from the BabelNet semantic network.
- Then, by analysing the corresponding contextual information and comparing and contrasting it with the whole Wikipedia corpus, a vectorial representation of the given synset is obtained.

two sorts of vectors

- On the basis of lexical specificity two types of representations are built: lexical and unified.
  - The *lexical* vector representation *lex*$_c$ of a concept c has lemmas as its individual dimensions.
  - The *unified representation* has concepts as individual dimensions.

Unified representation

- The algorithm first clusters together those words that have a sense sharing the same hypernym according to the BabelNet taxonomy
- Next, the specificity is computed for the set of all the hyponyms, even those that do not appear in the sub-corpus SC_c ;
- The binding of a set of sibling words into a single cluster represented by their common hypernym
  - transforms the representations to a unified semantic space;
  - allows to see the clustering as an implicit disambiguation process.



Crane (bird)			Crane (machine)		
English	French	German	English	French	German
shore_bird$^{1}_{n}$	‡famille_des_oiseaux$^{1}_{n}$	‡vogel-familie$^{1}_{n}$	∗lifting device$^{1}_{n}$	∗dispositif de levage$^{1}_{n}$	∗hebevorrichtung$^{1}_{n}$
bird$^{1}_{n}$	∗limicole$^{1}_{n}$	∗charadrii$^{1}_{n}$	‡construction$^{4}_{n}$	navire$^{1}_{n}$	radfahrzeug$^{1}_{n}$
∗wading_bird$^{1}_{n}$	oiseau_aquatique$^{2}_{n}$	†vogel_gattung$^{1}_{n}$	platform$^{1}_{n}$	limicole$^{1}_{n}$	†lenkfahrzeug$^{1}_{n}$
oscine_bird$^{1}_{n}$	tollé$^{2}_{n}$	wirbeltiere$^{2}_{n}$	warship$^{1}_{n}$	◇vaisseau$^{2}_{n}$	regler$^{3}_{n}$
†bird_genus$^{1}_{n}$	gallinacé$^{1}_{n}$	fleisch$^{1}_{n}$	electric circuit$^{1}_{n}$	spationef$^{1}_{n}$	reisebus$^{1}_{n}$
‡bird_family$^{1}_{n}$	◇classe$^{1}_{n}$	tier um$^{1}_{n}$	◇vessel$^{2}_{n}$	‡construction$^{2}_{n}$	charadrii$^{1}_{n}$
◇taxonomic_group$^{1}_{n}$	occurence$^{1}_{n}$	reiher$^{1}_{n}$	boat$^{1}_{n}$	†véhicule$^{3}_{n}$	güterwagen$^{2}_{n}$

word$^{p}_{n}$ is the p^{th} sense of the word with part of speech *n*.

Word senses marked with the same symbol across languages correspond to the same BabelNet synset.

Set of concepts associated to words

- Given these representations for individual word senses, the goal is to associate the set of concepts, i.e., BabelNet synsets, $C_w = \{c_1, ..., c_n\}$ with a given word *w*.
  - If *w* exists in the BabelNet dictionary, the set of associated senses of the word can be obtained as defined in the BabelNet sense inventory.
  - Use of *piped links*. Piped link is a hyperlink appearing in the body of a Wikipedia article, providing a link to another Wikipedia article, such as [[dockside_crane|Crane_(machine)]] is a hyperlink that appears as *dockside_crane* in the text, but takes the user to the Wikipedia page titled Crane_(machine).
    - In so doing, a set of concepts for the words not covered by BabelNet can be obtained.

Application: Semantic Similarity

- Once we have the set *C_w* of concepts associated with each word *w*, we first retrieve the set of corresponding unified vector representations.

- Then, the square-rooted Weighted Overlap (Pilehvar et al., 2013) as vector comparison method can be used,

$$WO(v_1, v_2) = \frac{\sum_{q \in O} \left(rank(q, v_1) + rank(q, v_2) \right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$

where *O* is the set of overlapping dimensions between the two vectors and *rank(q, v_i)* is the rank of dimension *q* in the vector *v_i*.
The similarity between words *w₁* and *w₂* is computed as the similarity of their closest senses

$$sim(w_1, w_2) = \max_{v_1 \in \mathcal{C}_{w_1}, v_2 \in \mathcal{C}_{w_2}} \sqrt{WO(v_1, v_2)}$$

NASARI taste & see

Each line in the vectors corresponds to a BabelNet synset. In the cases where the BabelNet synset is associated with a Wikipedia page, the Wikipedia page title is written in the second column. Otherwise it is written -NA-.

1. [Lexical vectors] the dimensions correspond to lemmas.

Files: NASARI_lexical_*.txt

Format (TAB separated):

```
BabelSynsetId  WikipediaPageTitle  lemma1_weight1  lemma2_weight2 ...
```

2. [Unified vectors] the dimensions correspond to the BabelNet synsets.

Files: NASARI_unified_*.txt

Format (TAB separated):

```
BabelSynsetId  WikipediaPageTitle  synset1_weight1 synset2_weight2 ...
```

The dimensions of lexical and unified vectors are separated from the weights using an underscore. Also, the vectors are truncated to the non-zero dimensions only and sorted according to the weights of their dimensions.

NASARI taste & see

```
3. [Embed vectors] are embedded vector representations of 300 dimensions:
  Files: NASARI_embed_*.txt
  Format (SPACE separated):
    BabelSynsetId  WikipediaPageTitle  dimension1  dimension2 ...
dimension300

Continuous vector representations (NASARI_embed) of BabelNet synsets
constructed by combining lexical vectors and the pre-trained models of
Word2Vec (300 dimensions).
```

I vettori di NASARI sono disponibili all'URL
http://lcl.uniroma1.it/nasari/

automatic summarization

di.unito.it
DIPARTIMENTO
DI INFORMATICA

credits

- E. Hovy, Chapter *Text Summarization*, in R. Mitkov (Ed.), The Oxford handbook of computational linguistics, Oxford University Press, 2005
- D. Jurafsky and J. H. Martin, *SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2009
- Eduard Hovy and Daniel Marcu, ACL Tutorial on Text Summarization, ACL 1998, Université de Montréal Montréal, Québec, Canada

a definition

- The goal of text summarization is to produce an abridged version of a text which contains the important or relevant information.
  - an abstract of a scientific article, a summary of email threads, a headline for a news article, or the short snippets returned by web search engines to describe each retrieved document.

goals

- Indicative: give an idea of what is there, provides a reference function for selecting documents for more in-depth reading
- Informative: a substitute for the entire document, covers all the salient information in the source at some level of detail
- Critical: evaluates the subject matter of the source, expressing the abstractor's view on the quality of the work of the author

kinds of automatic summarization

- Extracts are summaries created by reusing portions (words, sentences, etc.) of the input text verbatim, while
- Abstracts are created by re-generating the extracted content
  - Paraphrase, generation

kinds of automatic summarization

- Output: User-focused (or topic-focused or query focused): summaries that are tailored to the requirements of a particular user or group of users
- Background: Does the reader have the needed prior knowledge?
  - Expert reader vs. Novice reader
- General: summaries aimed at a particular –usually broad – readership community

Summarisation approaches

- Shallow approaches
  - Syntactic level at most
  - Typically produce extracts
  - Extract salient parts of the source text and then arrange and present them in some effective manner
- Deeper approaches
  - Sentential semantic level
  - Produce abstracts and the synthesis phase involves natural language generation.
  - Knowledge-intensive, may require some domain specific coding

single doc versus multiple doc summarisation

- In single document summarisation we are given a single document and produce a summary.
  - Single document summarisation is thus used in situations like producing a headline or an outline, where the final goal is to characterise the content of a single document.
- In multiple document summarisation, the input is a group of documents, and our goal is to produce a condensation of the content of the entire group.
  - We might use multiple document summarisation when we are summarising a series of news stories on the same event, or whenever we have web content on the same topic that we'd like to synthesise and condense.

parameters

- Compression rate (summary length/source length)
- Audience (user-focused vs. generic)
- Relation to source (extract vs. abstract)
- Function (indicative vs. informative vs. critical)
- Coherence: the way the parts of the text gather together to form an integrated whole
  - Coherent vs. incoherent
  - Incoherent: unresolved anaphors, gaps in the reasoning, sentences which repeat the same or similar meaning (redundancy) a lack of organisation

approaches comparison

- NLP/IE:
  - Approach: try to 'understand' text—re-represent content using 'deeper' notation; then manipulate that.
  - Need: rules for text analysis and manipulation, at all levels.
  - Strengths: higher quality; supports abstracting.
  - Weaknesses: speed; still needs to scale up to robust open-domain summarisation.

- IR/Statistics:
  - Approach: operate at lexical level—use word frequency, collocation counts, etc.
  - Need: large amounts of text.
  - Strengths: robust; good for query-oriented summaries.
  - Weaknesses: lower quality; inability to manipulate information at abstract levels.

relevance criteria

Position in the text

- Important sentences occur in specific positions
  - *"lead-based" summary* (just take first sentence(s)!)
  - Important information occurs in specific sections of the document (introduction/conclusion)
  - Experiments:
    - In 85% of 200 individual paragraphs the topic sentences occurred in initial position and in 7% in final position

Title method

- Title of document indicates its content
  - Not true for novels usually
  - What about blogs …?
- Words in title help find relevant content
  - Create a list of title words, remove "stop words"
  - Use those as keywords in order to find important sentences

Optimum Position Policy (OPP)

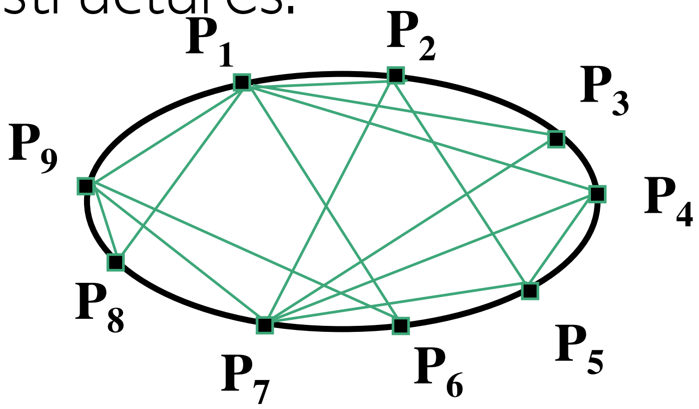
- Relevant sentences are located at positions that are genre-dependent; these positions can be either known or determined automatically through training
  - Step 1: For each article, determine the overlap between sentences and the index terms (e.g., title terms)
  - Step 2: Determine a partial ordering over the locations where sentences containing important words occur: Optimal Position Policy (OPP)

Cue phrases method

- Important sentences contain cue words/indicative phrases,
  - "The main aim of the present paper is to describe…"
  - "The purpose of this article is to review…"
  - "In this report, we outline…"
  - "Our investigation has shown that…"
- Some words are considered bonus others stigma
  - bonus: comparatives, superlatives, conclusive expressions, etc.
  - stigma: negatives, pronouns, *etc.* non-important sentences contain 'stigma phrases' such as hardly and impossible.
- These phrases can be detected automatically
- Method: Add to sentence score if it contains a bonus phrase, penalise if it contains a stigma phrase.

Cohesion-based methods

- Important sentences/paragraphs are the highest connected entities in more or less elaborate semantic structures.



- Classes of approaches
  - word co-occurrences;
  - local salience and grammatical relations;
  - co-reference;
  - lexical similarity (WordNet, lexical chains);
  - combinations of the above.

Cohesion: word co-occurrence

- Apply IR methods at the document level: texts are collections of paragraphs
  - Use a traditional, IR-based, word similarity measure to determine for each paragraph P_i the set S_i of paragraphs that P_i is related to.
- Method:
  - determine relatedness score S_i for each paragraph,
  - extract paragraphs with largest S_i scores.

3 (to 1) steps

- Text summarisation systems are generally described by their solutions to the following three problems:

- *Content Selection*: What information to select from the document(s) we are summarising. We usually make the simplifying assumption that the granularity of extraction is the sentence or clause. Content selection thus mainly consists of choosing which sentences or clauses to extract into the summary.
- *Information Ordering*: How to order and structure the extracted units.
- *Sentence Realisation*: What kind of clean up to perform on the extracted units so they are fluent in their new context.

unsupervised algorithm

- The simplest unsupervised algorithm is to select sentences that have more salient or informative words.
  - Sentences that contain more informative words tend to be more extract-worthy.
- *Saliency* is usually defined by computing the topic signature, a set of salient or signature terms, each of whose saliency scores is greater than some threshold θ.
  - Saliency could be measured in terms of simple word frequency, but frequency has the problem that a word might have a high probability in English in general but not be particularly topical to a particular document.
- Lexical specificity can thus be adopted in order to individuate the most salient terms, and to score the sentences where they appear.

a simple *extractive* algorithm

- reduce the document size of e.g., 10%, 20%, 30%

1. individuate the topic of the text being summarised; the topic can be referred to as a (set of) NASARI vector(s):

   v_{t1} = {*term*$_1$*_score*, *term*$_2$*_score*, …, *term*$_{10}$*_score* }
   v_{t2} = {*term*$_1$*_score*, *term*$_2$*_score*, …, *term*$_{10}$*_score* }
   …

2. create the context, by collecting the vectors of terms herein (this step can be repeated, by dumping the contribution of the associated terms at each round);

3. retain paragraphs whose sentences contain the most salient terms, based on the Weighted Overlap, *WO*(v_1,v_2)

   - rerank paragraphs weight by applying at least one of the mentioned approaches (*title*, *cue*, *phrase*, *cohesion*).

NASARI (lexical) subset

- two distribution files are provided for NASARI, that require different resources allocation.
  - *dd-nasari.txt*. a subset of NASARI (obtained by truncating vectors at 10 features). 3,587,754 vectors, ~600MB;

    https://goo.gl/85BubW
  - *dd-small-nasari-15.txt*. a subset of NASARI. same filtering as above, with 15 features + intersection with 60K lemmas in the Corpus of Contemporary American English: 13,084 vectors, 2MB storage (many entities removed here…).
- the second one has been extracted for starting our experimentation; the second one is intended to explore the resource in a richer (though reduced) flavour.

documents for summarisation

- text documents are provided for summarisation purposes:
  - *Andy-Warhol.txt*
  - *Ebola-virus-disease.txt*
  - *Life-indoors.txt*
  - *Napoleon-wiki.txt*
  - *Trump-wall.txt*
- do experiment with different compression rates: 10%, 20% and 30%.

evaluation

- evaluation can be performed based on two complimentary metrics
  - BLEU (bilingual evaluation understudy) regarding precision; and
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as regards as recall.

BLUE (bilingual evaluation understudy)

- scoring function that has been worked out to assess systems for automatic translation
  - build a reference summary, as a list of relevant terms that should be present.
  - compare the set of terms in the automatic summary (which we call candidate summary,) to those in the candidate summary.
  - the BLEU score is computed as $P = m/w_t$ that is the fraction of terms from the candidate that are found in the reference, where m is the number of terms in the candidate that are in the reference, and w_t is the size of the candidate
- precision in IR is customarily defined as

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- This metrics estimates in how far the words (and/or n-grams) in the human reference summaries appeared in the summaries built by the system

  - ROUGE-N: Overlap of N-grams between candidate and reference summary.
  - ROUGE-1 refers to the overlap of unigram (each word) between the system and reference summaries.

- recall in IR is customarily defined as

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$