Examining Crosslingual Word Sense Disambiguation

Liling Tan

School of Humanities and Social Sciences

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Masters of Arts

2013

Acknowledgements

This thesis is made possible by the mentoring and wise advice of Dr. Francis Bond and the authors of the resources that were used to examine Crosslingual Word Sense Disambiguation (CLWSD). The author of this thesis thanks also the lecturers and colleagues from the Universität des Saarland for their impartation of knowledge during the 10 months Erasmus exchange and occasional advice during the thesis process. Finally the author's gratitude extends to Erasmus Mundus MULTI committee which had made the exchange possible.

Abbreviations

ConceptID (*Concept IDentification number*) refers to the unique 8 digit identification number followed by a dash and its Part-of-Speech (POS tag). E.g. ***07470671-n*** refers to the synset defined as “a formal contest in which two or more persons or teams compete”; aka as **SenseID** (*Sense IDentification number*).

CLWSD (*Crosslingual Word Sense Disambiguation*) is the computational task of correctly identifying the foreign language translations of a polysemous word given an English context sentence

LDA (*Latent Dirichlet Allocation*) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

LSA (*Latent Semantic Analysis*, aka. *Latent Semantic Indexing*) is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

NLP (*Natural Language Processing*) is the multidisciplinary field between computer science, artificial intelligence, cognitive science and linguistics concerned with the computational understanding and production of human (natural) language.

NMF (*Non-negative matrix factorization*) is a group of algorithms in multivariate analysis and linear algebra where a matrix, V, is factorized into two matrices W and H.

NTU-MC (*Nanyang Technological University - Multilingual Corpus*) is a multilingual corpus built by the computational linguistic group in NTU

OMW (*Open Multilingual Wordnet*) is a collection of open source wordnets in a variety of languages, all linked to the Princeton Wordnet of English (see http://www.casta-net.jp/~kuribayashi/multi/)

PMI (*Pointwise Mutual Information*) is a statistical associative measure of how much a word tells us about another word

SemEval (*Semantic Evaluation*) is an ongoing series of evaluation workshops on computational semantic analysis systems (see http://en.wikipedia.org/wiki/SemEval)

Tf-idf (*Term frequency-inverse document frequency*) is a matrix statistic that reflects the importance of a word to a document in a corpus.

Topical CLWSD (*Topic-model based CLWSD*) is the task of using topic-models to resolve lexical disambiguity

WN (*WordNet*) is a lexical database of words and their respective psycholinguistically motivated concepts

WSD (*Word Sense Disambiguation*) is the computational task of correctly identifying the sense of a polysemous word given a context sentence

XLING (*CrossLINGual*) is the software created for thesis to attempt the Topical CLWSD task

XLING_SnT (*CrossLINGual Similar and Translate*) is the baseline model of the XLING software described in chapter 6

XLING_TnT (*CrossLINGual Topicalize and Translate*) is the main model of the XLING software described in chapter 6

Summary

Understanding human language computationally remains a challenge at different levels, phonologically, syntactically and semantically. This thesis attempts to understand human language's ambiguity through the Word Sense Disambiguation (WSD) task. Word Sense Disambiguation (WSD) is the task of determining the correct sense of a word given a context sentence and topic models are statistical models of human language that can discover abstract topics given a collection of documents.

This thesis examines the WSD task in a crosslingual manner with the usage of topic models and parallel corpus. The thesis defines a *topical crosslingual WSD* (Topical CLWSD) task as two subtasks (i) **Match** and **Translate**: finding a match of the query sentence in a parallel corpus using topic models that provides the appropriate translation of the target polysemous word (ii) **Map**: mapping the word-translation pair to disambiguate the concept respectively of the *Open Multilingual WordNet*. The **XLING** WSD system has been built to attempt the topical WSD task. Although the **XLING** system underperforms in the topical WSD task, it serves as a pilot approach to crosslingual WSD in a knowledge-lean manner.

Other than the WSD task, the thesis briefly presents updates on the ongoing work to compile multilingual data for the Nanyang Technological University-Multilingual Corpus (NTU-MC). Both the NTU-MC project and the XLING system are related in their attempts to build crosslingual language technologies.

The rest of the thesis will be as follows:

- **Chapter 1** provides an introduction and motivation to the Topical WSD task.
- **Chapter 2** briefly surveys the different representations of meaning and concludes with the thesis' take on meaning
- **Chapter 3** provides an overview of different WSD evaluation methods and approaches and presents the Topical match, translate and map approach to WSD
- **Chapter 4** reviews the available knowledge sources for WSD and highlights the resources used for the Topical WSD task
- **Chapter 5** summaries topic modeling and its relation to Natural Language Processing as well as its usage in the Topical WSD task
- **Chapter 6** describes, evaluates and discusses the Topical CLWSD task
- **Chapter 7** concludes the main thesis and discusses future work.
- **Chapter 8** presents updates on the ongoing NTU-MC project, a parallel corpora project that may be used for future work on CLWSD.

Examining Crosslingual Word Sense Disambiguation

Liling Tan

Contents

Chapter 1

Introduction

Language is ambiguous by nature. Consider the word *match* in the following sentences:

(a) Mustafa may not be as fancy as some of Singapore's other malls, but it has a great range of items, and good prices to *match*.

(b) So if you hanker for watching some high octane local blade action, call the NIHL at (65) 6276 0364 for more *match* schedules and updates.

The occurrences of the word *match* in the two sentences clearly denote different meanings; respectively, they mean :[1]

(i) *be compatible, similar or consistent*

(ii) *a formal contest in which two or more persons or teams compete*

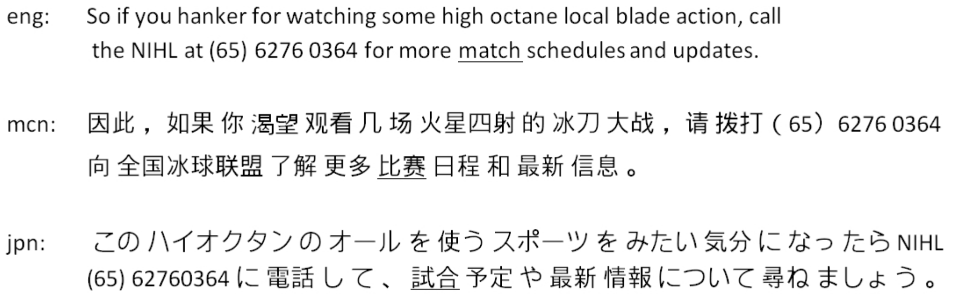
Resolving such lexical ambiguities computationally is an Artificial Intelligence-complete (AI-complete) problem for machines (Mallery, 1988). The computational linguistic task to identify the correct meaning of words in the given context is called Word Sense Disambiguation (WSD).

1.1 The Power of Parallel

Consider a multilingual approach to WSD where translations of a polysemous word provide complementary information on a different range of meanings (i.e. every translation of a polysemous word encodes a different set of

[1] The meanings are taken from the Princeton English WordNet (Fellbaum, 1998).

pycholinguistic concepts). Different translations provide complementary information that is synergistic in reducing word ambiguity. For example the sentence (b) is translated as such in Chinese and Japanese:



eng: So if you hanker for watching some high octane local blade action, call the NIHL at (65) 6276 0364 for more match schedules and updates.

mcn: 因此，如果 你 渴望 观看 几 场 火星四射 的 冰刀 大战，请 拨打（65）6276 0364 向 全国冰球联盟 了解 更多 比赛 日程 和 最新 信息 。

jpn: この ハイオクタン の オール を 使う スポーツ を みたい 気分 に なっ たら NIHL (65) 62760364 に 電話 し て 、 試合 予定 や 最新 情報 について 尋ね ましょ う 。

Given the respective sense inventory in English, Chinese and Japanese,[2] we consider all the possible senses of the word *match* and its translations. We can easily disambiguate the sense in sentence (b) as ***07470671-n***[3]: "a formal contest in which two or more persons or teams compete".

eng: {00041188-n, 00456199-n, 00457382-n, 00557588-n, 07456188-n, 07458453-n, 07464725-n, 07468116-n, **07470671-n**, 07472327-n, 07472657-n}

mcn: {03728437-n, 03728811-n, 03728982-n, 05696020-n, **07470671-n**, 07988857-n, 09626238-n, 09900981-n, 13596673-n}

jpn: {00446493-n, 00456199-n, 01168961-n, 07456188-n, **07470671-n**, 13596235-n}

Navigli & Ponzetto (2012a) validated the effective use of multilingual information on disambiguating word senses using an ontological Knowledge Base (KB) through graph-based WSD methods. Their experiments were carried out in 6 closely related European languages. They achieved 6% improvement on SemEval-2010's all-words WSD evaluation and an improvement of 2% in evaluating SemEval-2010's lexical substitution task. They concluded with the paper titled "*Joining Forces pays off*".

This thesis' direction is in the same exploration of crosslingual information sense disambiguation but in a knowledge-leaner approach (using topic

[2] i.e. Princeton WordNet, Chinese WordNet (Xu et al., 2008) and Japanese WordNet (Isahara et al., 2008)

[3] In the WordNet, each sense/concept is given an 8 digit identification number, ***conceptID/senseID*** followed by a dash and its Part-of-Speech (POS) tag, e.g. ***-n*** for noun

models). WSD tasks are normally defined in a monolingual context, where systems provide the appropriate sense of a polysemous word given an English sentence. Hence we explore the possibility of finding the closest match of the context English sentence with sentences from a parallel corpora and consequently finding the correct sense through the polysemous word and its translations. We call the task *Topical Crosslingual Word Sense Disambiguation* (Topical CLWSD).

1.2 Topical CLWSD

Previously, researchers attempted WSD through topic models by using topic models as global context features in supervised WSD tasks (e.g. Cai et al. (2007) ; Boyd-Graber et al. (2007)) and by incorporating topics as weights into probablistic WSD models (e.g. Li et al. (2010)). We propose an alternative sense disambiguation method using parallel corpus as a medium of external knowledge to reduce the sense ambiguity of polysemous word. We define the task of *Topical Crosslingual Word Sense Disambiguation* with two subtasks:

(1) Given a context sentence, find an equivalent sentence and its translation from a parallel corpus

(2) Using the word-alignment of the polysemous word from the matched sentence, generate the ambiguity reduced conceptID(s)

For example, given the query sentence, Q, :

Q: "*In contrast, the rambling valedictory press conference last fall of Notre Dame's football* ***coach****, Lou Holtz, was criticized by one sportswriter for its 'absence of any real sense of closure'.*"

We use Latent Dirichlet Allocation (LDA) to find the top ranking topic and from a parallel corpus and we find a matching English sentence that shares the same top ranking topic, M with its Spanish translation, T, e.g.:

M: "*He failed as a* ***coach*** *for the reason that other great players have failed as coaches: he thought about himself too much.*"

T: "*Fracasó como entrenador por la razón de que otros grandes jugadores han fallado como entrenadores.*"

By taking the word-alignment of ***coach*** from M and T, we check the word pair (***coach:entrenador***) in the Open Multilingual Wordnet (Bond and Pease, 2013) and the WSD system responds with the appropriate sense(s) that have ***coach*** in the English wordnet and ***entrenador*** in the Spanish wordnet; in this case the ***coach:entrenador*** word pair yields the conceptID ***09931640-n*** "someone in charge of training an athlete or a team".

To evaluate the first subtask of finding the correct translation for a polysemous word in the query sentence, we attempt the Crosslingual Word Sense Disambiguation (CLWSD) task in SemEval-2013. The aim of CLWSD task in SemEval-2013 is to evaluate systems that provide the appropriate translation(s) for a polysemous word given a context sentence. Thereafter, we map the gold answers and system answers to the Open Multilingual WordNet (OMW) and check the accuracy of our system in providing the correct conceptID using traditional precision, recall, F-score measures.

1.3 Thesis Walkthrough

The rest of the thesis is structured as such:

Chapter 2 briefly defines various notions of meaning from different fields, describes different approaches to lexical semantics and discusses the usage of latent topics as semantic knowledge for the *Topical CLWSD* task.

Chapter 3 surveys the different evaluation methods, different approaches to the WSD and asserts topic models as the thesis' approach to crosslingual WSD.

Chapter 4 gives an overview of notable resources for Natural Language Processing (NLP) and lists the resources used in the *Topical CLWSD.*

Chapter 5 provides a primer on topic models in NLP.

Chapter 6 describes, evaluates and discusses the Topical CLWSD task.

Chapter 7 concludes the main thesis and discusses future work.

Chapter 8 presents updates on the ongoing NTU-MC project, a parallel corpora project that may be used for future work on CLWSD.

Chapter 2

The Meanings of Meaning

Meaning is classically defined as having two components (Lyons, 1977):

1. ***Reference***, anything in the referential realm (i.e. anything, real or imagined, that a person may talk about) denoted by a word or expression, and

2. ***Sense***, the system of paradigmatic and syntagmatic relationships between a lexical unit and other lexical units in a language

The first section of this chapter briefly describes the various notions of meaning from different fields (section 2.1) and the following section gives an overview to the different approaches to lexical semantics representation (section 2.2). This chapter concludes with the discussion on representing semantic knowledge as latent topics and its relevance to the thesis task, viz. *Topical Crosslingual WSD.*

2.1 Different Meanings of Meaning

Formal Semantics

Traditionally, formal semantics understand meaning through the construction of precise mathematical/logical models that define relations between the linguistic expressions (***symbols***) and the referential worlds (***references***; real or imagined).
Theories of formal semantics are non-psychological (truth-conditional, model-theoretic, possible worlds, situation, etc.), hence meanings can be *out of mind* yet *of the world.* For example, the word *dog* can refer to "canis familiaris" (let's call the lexeme $\boldsymbol{dog}_1$); regardless of the mind's ability to perceive $\boldsymbol{dog}_1$, it exist in the natural world nevertheless. Thus to communicate the meaning or idea of $\boldsymbol{dog}_1$, one will use the linguistic symbol, *dog.*

Cognitive Semantics

Cognitive semantics adds the dimension of mental ***concepts*** to the formal semantics diachotomic meaning (of *symbols* and *references*). Theories of cognitive semantics assume that there are psycholinguistic constructs stored in our long-term memory and the usage of these constructs interacts with the world to form meaning dynamically (Ungerer & Schmid, 1996).

Often, cognitive semantic studies (e.g. prototypes and categories, figure and grounding, frames and construction, etc.) focus on the mental conceptualization of the world and inter-concept mappings and undermine the linkage between the concepts and their linguistic symbols.

Lexical Semantics

Lexical semantics provides structures that systematically map linguistic expressions to cognitive concepts. Different from the cognitive semanticists, lexical semanticists are interested in the most optimal approach towards a unified ***label-concept*** system to the representation of meaning.

Previously we have discussed linguistic expressions/symbols as words, but the dogmatic term should be *label.* Other than a flat representation of a *word*, there are variants of the same word that refers to different ***concepts*** and possibly adhere to the different grammaticality (i.e. with different Parts-

Of-Speech). Another term that is frequently used in place of *label* is *token.*[1] The aim of the Word Sense Disambiguation task is to automatically identify the different label-concept mappings for polysemous labels.

2.2 Different Representations of Meaning

Textual semantic knowledge can be conceived as knowledge about relations between a *word*, its *concept(s)* and their analogous *percepts.* Different aspects of these relations have been studied:

- **Word-Concept(s)** relations: Knowledge of the word *dog* can refer to the concepts ***dog***$_1$: "a member of the genus Canis"; ***dog***$_2$:"a dull unattractive unpleasant girl or woman"; ***dog***$_3$:"a smooth-textured sausage of minced beef or pork usually smoked"

- **Concept-Concept(s)** relations: Knowledge that ***dog***$_1$ is a kind of ***animal***$_1$ (hypernym), and ***poodle***$_1$ (hyponym) is a kind of ***dog***$_1$, the ***tail***$_1$ (holonym) is part of the ***dog***$_1$.

- **Concept-Percepts** relations: Knowledge of how a dog sounds (***bark***$_1$), how a dog is different from a cat (***bark***$_1$ vs ***meow***$_1$), what dogs likes (the skeletal ***bone***$_1$), etc.

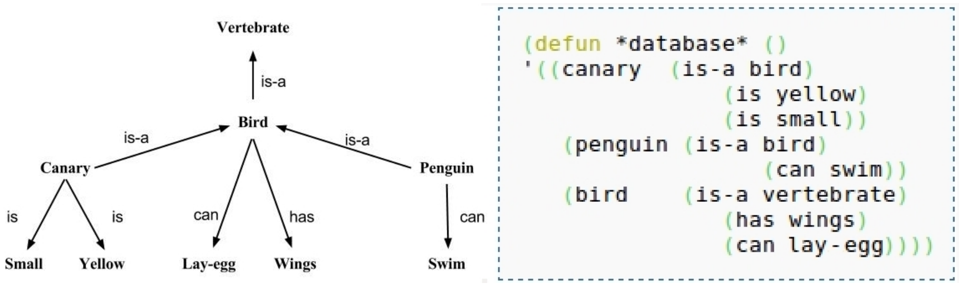
- **Word-Words** relations: the word *dog* tends to collocate with words such as *wag*, *pet*, *bone*, *cat*, *chilli*, *hot*, *slut*, *frump*, etc.

Different approaches to semantic modeling tend to focus on different aspects of relational knowledge. Computationally, textual semantic models can either be automatically learnt from natural texts or through manual annotations subscribing to certain ontological framework. The following subsections describe the different approaches to lexical semantics.

[1] For simplicity of reading, we will refer to labels as labels, words or tokens interchangeably for the rest of this thesis.

2.2.1 Semantics from Networks

Since the late 1960s, semantic knowledge is traditionally represented in a semantic network with *abstract propositions* as its edges and ***concepts*** as nodes. For example, ***canary*** *is-a* ***bird***; ***bird*** *has* ***wings***; ***penguin*** *can* ***swim***.



```
(defun *database* ()
'((canary  (is-a bird)
               (is yellow)
               (is small))
   (penguin (is-a bird)
                 (can swim))
   (bird    (is-a vertebrate)
               (has wings)
               (can lay-egg))))
```

Figure 2.1: Canary Example in a Semantic Net - Graphical representation (left) and in `Lisp` (right)

Semantics Network from Ontology

Early researches on semantic networks focused on *Concept-Concept(s)* relations where the distinction between words and concepts is typically collapsed under this approach and the abstract propositions were manually coded into a conceptual database (Collins & Quillian, 1969). Rogers & McClelland (2008) was able to automatically learnt these conceptual relations by looking at word frequencies however the scale of their experiment was rather small (∼40 nodes: 8 input concepts mapping to 30+ attributional concepts) and the propositions they attempted was quite rudimentary (3 edge types: *is*, *can*, *has*). Mapping large scale abstract conceptual relations remains unsolved. But in the edge of massive information databases (see section 4.2.2), extracting knowledge relations mapping has shifted from the traditional ontological classifications to Knowledge-Base Population tasks (Ji & Grishman, 2011).

Semantic Network from Natural Texts

The other approach of semantic representation focuses more on the associative relations between words in natural texts (*Word-Words* and *Word-Percepts* relations) and their *Word-Concepts* mappings. For instance, we generally associate the word *bird* with *wings*, *fly*, *chirp*, *nest*; but we also accept affiliating *bird* with less frequently associated words such as *turkey*, *thanksgiving*, *early*, *bee*.

These distributional expectations reflect that the polysemous nature of the word *bird* to refer to bird as a taxonomical category, to refer to bird as food and multiple idiomatic uses of bird. Psycholinguistically, semantic network suggests that semantic representation exists in form of a network and the retrieval of meaning is made through lexical access using our short-term memory. Computationally, viewing lexical meanings in form of a network is not unlike graphical abstract data structure and the retrieval of meaning can be formulated as finding the concept node with the most optimal traversal of the relation path. The automation of learning Word-Percepts and Word-Concepts relations from large-scale linguistic corpora was implemented in the late 1990s (e.g. Lund & Burgess (1996); Landauer & Dutnais (1997)).

Semantic Network Unified (WordNet and SUMO)

WordNets are large lexicons that encode psycholinguistically motivated concepts. These concepts are represented by sets of cognitive synonyms called synsets (a word entry in a thesaurus). In addition, each synset also has a gloss that explains the meaning of the concept in plain english (like a dictionary). The Princeton WordNet (Miller et al. (1990); Fellbaum (1998)) is the de facto semantic resource for WSD. The latest version of the Princeton WordNet 3.0 contains 117,000 synsets. Then Niles & Pease (2001a) began the alignment of WordNet concepts onto Suggested Upper Merged Ontology (SUMO) which unifies the ontological and natural text approaches to semantic representation. WordNet today is the idealized combination of both approaches where Word-Concepts/Percepts mappings are either manually input or automatically learnt from a corpus Concept-Concepts mappings are available with SUMO.

Recently researches have implemented creative ways to create WordNets in other languages by using translations of manually tagged English Corpora (e.g. Japanese WordNet: Bond et al., 2008; Diab & Resnik, 2002) or using ontology to build WordNets in other languages (e.g. EuroWordNet: Vossen (1998)) or creating new WordNets using monolingual or bilingual dictionaries (e.g. BalkaNet: Tufiş et al., 2004). Vice versa, the SUMO was enriched by the integration of concepts from different WordNets, Pease &

Fellbaum (2010) provided a brief history on the work of WordNet and SUMO integration.

2.2.2 Semantics as Vectors

Semantic knowledge can be thought as a two dimensional vector space where each word is represented as a point and semantic association is indicated by word proximity. The vector space model focuses on identifying the Word-Word and Word-Percept relations. The primary attraction to semantic vectors is that semantic knowledge can be automatically extracted with a raw corpus without manual annotations or lexicon building.

Researches adopted vector space model to Natural Language Processing tasks and achieved impressive results in emulating human language usage. Rapp (2003) used a vector-based model of word meaning and scored 92.5% on the Test of English as Foreign Language (TOEFL), where the average human score was 64.5%. Similarly, vector-based semantics scored 56% on multiple-choice analogy questions in the SAT college entrance test while humans score a 57% on average (Turney, 2006).

The vector space model was originally developed for the SMART information retrieval system where a point represents a document (instead of a word) and a query is represented as another point in the same space and relevance of documents is measured by their similarity (i.e. proximity) to the query vector (Salton, 1971). Instead of measuring document similarity, Deerwester et al. (1990) proposed word similarity measurement by representing documents as collections of words and the resultant vector of each document refers to a row vector in the term-document matrix. The general hypothesis of vector semantics assumes that statistical patterns of human word usage can be used to disambiguate word meanings (i.e. if two word vectors in any frequency based matrix are similar, they have similar meaning). For further readings on vector semantics, Turney & Pantel (2010) provided a survey on vector space models.

Vector-based WSD systems started with statistical tweaking of the vector matrix to achieve the state-of-art through normalization, weighting, discounting and smoothing (e.g. tf-idf: Singhal et al. (1996); PMI: Turney (2001)). Then studies attempted vector composition where the meaning of a target vector a in the context vector b is a function of the vectors, i.e. $c = f(a,b)$ (Kintsch & Kintsch (2001); McDonald & Brew (2004)). More recently, researches have injected syntactic compositions and additional knowledge (such as sense frequencies and ontological relations) to the vector composition, redefining the general class of vector models as: $c = f(a,b,R,K)$ where R = syntactic relations, and K = external knowledge (Mitchell & Lapata (2008); Erk & Padó (2008); Thater et al. (2011)).

2.2.3 Semantics from Latent Dimensions

Latent dimensions can be thought of as the subconscious workings of the lexical access hidden in the different usage of different words. Latent Semantic Analysis (LSA) and topic modeling are two prominent vectorial and statistical approaches to dimensionality reduction that emulates human's semantic access.

Landauer & Dutnais (1997) proved that LSA is able to emulate human's psycholinguistic behavior and Griffiths et al. (2007) had showed that topic models are capable of various human linguistic behavior such as inducing a perceptual ontological hierarchy and imitating human's semantic memory. Computationally, latent variables/spaces semantics have been successful in Word Sense Disambiguation tasks (e.g. Katz & Goldsmith-Pinkham (2006); Li et al. (2010); de Cruys & Apidianaki (2011)).

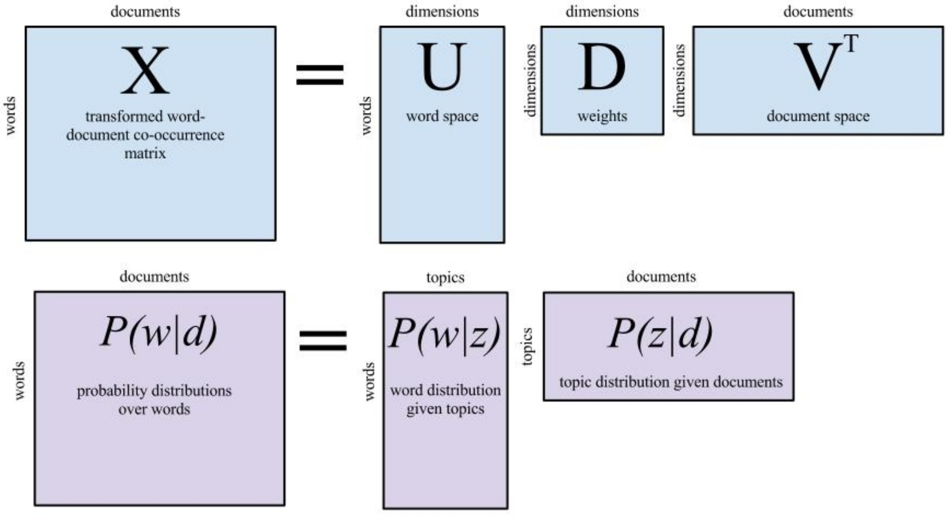


Figure 2.2: Latent Semantic Analysis (LSA) (top) and Topic Modeling(bottom)

Semantics in Latent Space

Latent Semantic Analysis (LSA) extracts spatial representation of words from a corpus of multiple sentences/documents. By feeding a word-document co-occurrence matrix to the LSA system, LSA decomposes the matrix into three smaller matrices, *U, D* and *V* (see figure 2.2):

- *U* provides an orthonormal basis for spatial representation of words (i.e. matches each word to a latent dimension)
- *D* weighs the dimensions (i.e. determines how different each dimension is different from each other)
- *V* provides an orthonormal basis for spatial representation of documents (i.e. matches each document to a latent dimension)

To measure word association, the cosine of the angle between the rows of U matrix has proven to be effective Landauer & Dutnais (1997). By reducing the dimensionality (i.e. the number of columns in U), statistical noise will decrease and latent correlations among words surfaces. Dimensionality reduction is achieved by the Singular Vector Decomposition (SVD) the term-document matrix.

Till et al. (1988) studied the time used to processing word meaning using a priming study where the participants read sentences that contains ambiguous words and then they were asked to choose perceptual words related to the sentence after varying delay times. For example, the sentence " *Thinking of the amount of garlic in his dinner, the guest asked for a mint*." and the lexical choices are {*money, candy, breath, coins*} .[2] They found that longer delay time primes more accurate choices. Landauer & Dutnais (1997) suggested that the same priming effect can be explained using LSA, the short delay could be represented by taking the cosine of the just the ambiguous word to the lexical choices and the long delay can be modeled when taking the cosine of the entire sentence to the lexical choices. Griffiths et al. (2007) argues that, psychologically, LSA can classify words into clusters and more critically find words that lie between two clusters and identify words that can appear in two clusters which are less useful in discriminating polysemous words.

[2] correct choices are underlined

Katz & Goldsmith-Pinkham (2006) implemented LSA to WSD by adding LSA reduced dimensions to K-nearest neighbor cosine similarity classifier to disambiguate word senses. They found that LSA required more unique contextual tokens to better differentiate senses. But their pure LSA classification method did not perform better than Term Frequency - Inverse Document Frequency (tf-idf) based classifier. Also, they improved the tf-idf classifier by merging LSA classification through a voting system used in Wicentowski et al. (2004).

de Cruys & Apidianaki (2011) explored the use of Non-Negative Matrix Factorization (NMF) to induce senses from latent factors using the surface words, ngrams and dependency-based context features. NMF also preforms dimensionality reduction like LSA but it applies a different factorization technique that uses Kullback-Leibler (KL) divergence instead of the Euclidean distance as in LSA. Minimizing the KL divergence is more representative of language phenomena than Euclidean distance (ED) because ED requires a normally distributed matrix but KL divergence allows Zipf (1949) skewed data. Also mathematically, NMF ensures non-negative probabilities which can be integrated into other NLP systems more easily; it is done by allowing only additive and non-subtractive relations. By mapping the centroid of the induced senses to a sense inventory, de Cruys & Apidianaki (2011) achieved better state-of-art results on SemEval-2010's WSI/WSD evaluation task. Psycholinguistically, NMF provides the same disambiguating information as LSA where *Word-Percept* relations are discovered by using latent correlations among words from matrix factorization.

Semantics in Latent Topics

Topic Modeling is a statistical approach to semantic representation that assumes hidden topics are embedded in a corpus. Topic modelling uses the same word-document co-occurrence matrix as LSA and NMF but the method to extract latent semantic knowledge is different. Different from LSA's vector decomposition and deletion, dimensionality reduction for topic modeling is achieved from statistical inference.

Topic model assumes the existence of latent topic variables that represent the gists of any set of correlated words. The probability distribution of a word (w) over each documents (d) is approximated by the probability distribution of the topics given the documents (aka the *gist* (g); where each topic is a probability distribution over words and each document is a probability distribution over topics.

Psycholinguistically in semantic intrusion studies, topic model could emulate the gist-based memory word association studies in the Deese-Roediger-McDermott (DRM) paradigm (Deese (1959); Roediger & Mcdermott (1995)). In the DRM paradigm, participants are introduced to a list of percepts {*bed, rest, awake, tired, dream, snooze, slumber, snore, nap, yawn*} that are associatively related to a word but it is not in the list (i.e. the lure word, *sleep*). Then participants are asked to recall all the words and 61% of the subjects falsely recalled *sleep*. Using topic model inference, Griffiths et al. (2007) reproduced the Roediger et al. (2001) DRM recall task and found that their topic inferred list of words from a corpus wrongly listed the lure word as human would. And the rank-order correlation of the lure word was at 0.437 at 95% confidence interval.[3]

The ability to emulate human's perception of sematically associative words (i.e. psycholinguistically mimetic) and the generative nature of topic modellings (i.e. computationally minimalistic) makes it an appealing method to find matching sentences for the Topical CLWSD task.

2.3 Thesis Take on Meaning

This thesis adopts the lexical approach to meaning encapsulated in pyscholinguistically *concepts* as defined in the WordNet and approaches the topical CLWSD task using the topic model's representation of meaning hidden within *latent topics*.

Although representing semantics with statistical inference may seem heretic, statistical inference on human behavioral tasks are psychologically grounded (Anderson & Schooler, 1991); more specifically topic models have shown success in capturing *Concept-Percepts* relations (chapter 3.2.3.2). Technologically, the probabilistic nature of the model allows it to be extensible with extra semantic knowledge and also easily integrated into other Natural Language Processing tasks. Using topic models, we craft the *Topical Crosslingual WSD* task where we use topic models and parallel dictionary entries to disambiguate meaning of words given a context sentence.

[3] Further discussion on topic modeling and its usage in WSD is found in chapter 3.2.3.2 and chapter 5

Chapter 3

Word Sense Disambiguation

Word Sense Disambiguation (WSD) is the ability to computationally determine the correct sense of a word given a particular context. The main aim of the Word Sense Disambiguation task is to correctly assign the labels to polysemous words given a context sentence. For example given the sentence "*The dog barks at the cat*", WSD systems should correctly label *bark* as ***bark***$_1$ ("*Woof!Woof!*", the onomatopoeia) as the correct sense in the sentence not tree ***bark***$_2$.

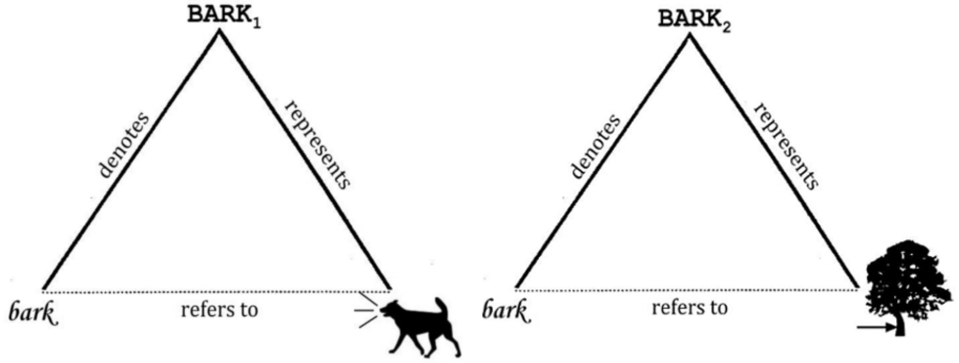


Figure 3.1: Polysemous `BARK`

This chapter of the thesis provides an overview of different evaluation methods (section 3.1) and approaches to the WSD task (section 3.2). And finally concludes with the Topic Models usage for WSD and brief concluding words on how topic models will be used in the *Topical Crosslingual WSD* task for this thesis.

3.1 Different WSD Evaluation Methods

As language technology evolves, the Word Sense Disambiguation (WSD) task grows in different flavors towards various research directions and for more languages:[1]

- ***Classic monolingual WSD*** evaluation task uses WordNet as its sense inventory and is largely based on supervised/semi-supervised classification with the manually sense annotated corpora:
  - *Classic English WSD* uses the Princeton WordNet as it sense inventory and the primary classification input is normally based on the SemCor corpus.
  - *Classical WSD for other languages* uses their respective WordNet as sense inventories and sense annotated corpora tagged in their respective languages. Often researchers will also tap on the SemCor corpus and aligned bitexts with English as its source language
- ***Multilingual WSD*** evaluation task focuses on WSD across 2 or more languages simultaneously, using their respective WordNets as sense inventories. It evolved from the Translation WSD evaluation task that took place in Senseval-2. A popular approach to multilingual WSD is to carry out monolingual WSD and then map the source language senses onto the corresponding translation of the target polysemous word.
- ***Crosslingual WSD*** (CLWSD) evaluation task is also focused on WSD across 2 or more languages simultaneously. Unlike the Multilingual WSD tasks, instead disambiguating senses by providing the concepts from a predefined sense inventory, systems participating in the CLWSD task provides the most appropriate translations of the target polysemous words in their respective target languages.
- ***Word Sense Induction and Disambiguation*** task is a combined task evaluation where the sense inventory is first induced from a fixed training set data, consisting of polysemous words and the sentence

[1] This section of the thesis was made available in Wikipedia `http://en.wikipedia.org/wiki/Word-sense_disambiguation#Task_design_choices` before the present submission

that they occurred in, then WSD is performed on a different testing data set.

3.1.1 Multilingual vs Crosslingual WSD

Multilingual and Crosslingual Word Sense Disambiguation (WSD) evaluation tasks focused on WSD across two or more languages simultaneously. While the Multilingual WSD evaluation task uses a fixed sense inventory (i.e. BabelNet), the sense inventory for the Crosslingual WSD evaluation task is built up on the basis of parallel corpora, e.g. the Europarl corpus.

Multilingual WSD

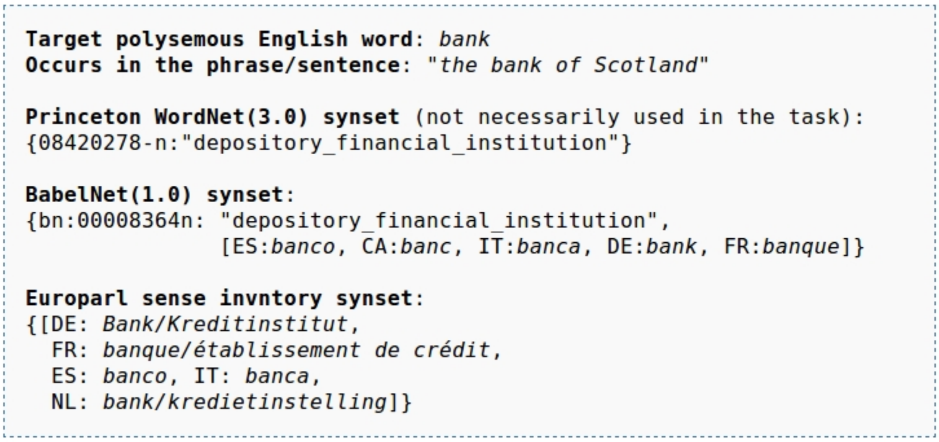
The Multilingual WSD task is introduced for the current SemEval-2013 workshop. The task is aimed at evaluating Word Sense Disambiguation systems in a multilingual scenario using BabelNet as its sense inventory. Unlike similar tasks, like *Crosslingual WSD* or the *Multilingual Lexical Substitution* where no fixed sense inventory is specified, Multilingual WSD uses the BabelNet as its sense inventory. Prior to the development of BabelNet, a bilingual lexical sample WSD evaluation task was carried out in SemEval-2007 on Chinese-English bitexts (Jin et al., 2007).

The multilingual WSD task follows the all-word version of classic WSD, where participating systems are expected to link all occurrences of noun phrases within arbitrary texts in different languages to their corresponding Babel synsets (Navigli & Ponzetto, 2012b). The evaluation criterion for the multilingual WSD task follows the standard precision, recall and F1 measures similar to the evaluation for classic WSD.

BabelNet

BabelNet is a very large multilingual semantic network with millions of concepts obtained from:

- an integration of WordNet and Wikipedia based on an automatic mapping algorithm and
- translations of the concepts (i.e. English Wikipedia pages and WordNet synsets) based on Wikipedia cross-language links and the output of a machine translation system



```
Target polysemous English word: bank
Occurs in the phrase/sentence: "the bank of Scotland"

Princeton WordNet(3.0) synset (not necessarily used in the task):
{08420278-n:"depository_financial_institution"}

BabelNet(1.0) synset:
{bn:00008364n: "depository_financial_institution",
               [ES:banco, CA:banc, IT:banca, DE:bank, FR:banque]}

Europarl sense invntory synset:
{[DE: Bank/Kreditinstitut,
  FR: banque/établissement de crédit,
  ES: banco, IT: banca,
  NL: bank/kredietinstelling]}
```

Figure 3.2: Example of a sense label in BabelNet and CLWSD task

Crosslingual WSD

The Crosslingual WSD (CLWSD) task is introduced in the SemEval-2007 evaluation workshop and re-proposed in SemEval-2010 as well as the current SemEval-2013 workshop. To facilitate the ease of integrating WSD systems into other Natural Language Processing (NLP) applications, such as Machine Translation and multilingual Information Retrieval, the crosslingual WSD evaluation task was introduced a language-independent and knowledge-lean approach to WSD.

The task is an unsupervised Word Sense Disambiguation task for English nouns by means of parallel corpora. It follows the lexical-sample variant of the Classic WSD task, restricted to 20 polysemous nouns. The evaluation criterion uses a weighted version of the precision and recall metric inspired by the English lexical substitution task in SemEval-2010. Participating systems in this evaluation task are free to use any corpus to build up their sense inventories (e.g. see figure 3.2).

3.2 Approaches to Word Sense Disambiguation

Navigli (2009) visualized WSD approaches as on a bidimensional space, where the vertical axis represents the ratio of sense-annotated to unlabeled data needed which determines the degree of machine learning supervision. The horizontal axis represents the amount of knowledge (e.g. lexical inventory, dictionaries, ontology, domain labels). The general WSD approaches can be summarized as points on the bidimensional space:

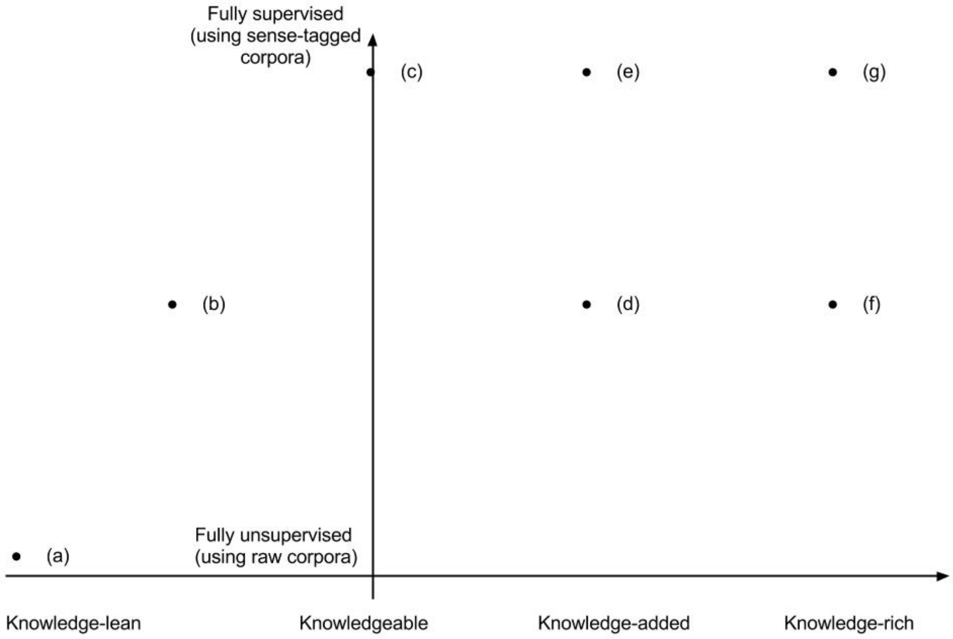

Fully supervised (using sense-tagged corpora)
(c)
(e)
(g)
(b)
(d)
(f)
Fully unsupervised (using raw corpora)
(a)
Knowledge-lean
Knowledgeable
Knowledge-added
Knowledge-rich

Figure 3.3: Different Approaches to WSD

(a) Fully unsupervised, knowledge-lean systems that do not use any amount of knowledge (often not even a sense-inventory and only using sentence-aligned parallel corpora)

(b) Minimally or semi-supervised systems (e.g. self-training or co-training) that uses little amount of sense-tagged data

(c) Fully supervised, knowledgeable systems (machine-learning classifiers) uses feature based machine methods with as much sense-tagged data as available (d,e) Knowledge-added systems that often exploit unstructured semantic knowledge (see next chapter) about corpus data such as sense dominance (i.e. sense frequencies), domain-labels, collocation lists, syntactic preferences, etc.

(f,g) Knowledge-rich systems uses structured knowledge sources such as ontological graph searches, gloss overlaps using dictionaries/thesauri

This rest of this section presents several methodologies that describe knowledge-rich systems, knowledge-added systems and knowledge-lean systems in Crosslingual WSD (CLWSD) context.

3.2.1 Knowledge-rich systems

Overlapping Senses

Knowledge rich approaches such as the Lesk algorithm (Lesk, 1986), Conceptual Density (Agirre & Rigau, 1996) and Random Walk algorithm (Mihalcea, 2006) are primarily overlap based system comparing the target polysemous word, its context, the corresponding synsets' glosses and semantic relations overlaps within a sense inventory/ontology. They suffer from data sparsity when there are no surface token matches between the queried data and training data. Mahapatra et al. (2010) overcame the overlapping sparsity issue by combining an overlapping measure with WordNet based lch similarity measure (Leacock and Chodorow, 1998) that accounted for semantic generalization. Their multi/crosslingual disambiguation was done first by matching the query sentence to the k-Nearest Neighbor in the training data ranked by their combined scoring function, then looking at the translation of the matched sentence.[2]

3.2.2 Knowledge-added systems

Most Frequent Sense

Zipf (1949) law established the relationship between the probability of language usage and word frequency (i.e. the frequency of any word is inversely proportional to its rank in the frequency table) and that lexical

[2] Mahapatra et al. (2010) OWNS system participated only in the French CLWSD task in SemEval-2010

access is based on principle of least effort. As much as the highly favored Most Frequent Sense (MFS) baseline produces rather high accuracy in WSD tasks, the MFS phenomenon is psycholinguistically motivated too. Balota & Chumbley (1984) concluded that high frequency words were named more easily in lexical decision task.

The variant of the MFS in the Crosslingual WSD task is the Most Frequent Translation (MFT); MFTs are the most frequent lemmatized translation result from the automated word alignment process (GIZA++).

Corpus-based Word Experts

For the CLWSD task in SemEval-2010, van Gompel (2010) UvT-WSD (Universiteit van Tilburg-WSD) system took the Tilburg Memory-Based Learner (TiMBL) classification approach sense disambiguation and selecting the translation using K-Nearest Neighbor. His system encoded corpus-based word experts (Ng & Lee (1996a); Hoste et al. (2002)) in form of global context features for the TiMBL classifier; local features includes words, n-grams, POS and lemma. All words from The UvT system was the top ranking system for Spanish and Dutch CLWSD (nld: 17.7%; spa: 23.42% precision), even with only the word expert feature it ranked second (nld: 15.93%, spa: 19.92%).

3.2.3 Knowledge-lean systems

Unsupervised Graphical Search

From a graphical approach to WSD, Silberer & Ponzetto (2010) attempted unsupervised crosslingual WSD using multilingual co-occurrence graphs from the adapted PageRank algorithm (Agirre & Soroa, 2009) and disambiguate the sentences by selecting the highest scoring vector on the computed Minimum Spanning Tree, similar to the Hyperlax algorithm (Véronis, 2004). Siberer and Ponzetto extended the multilinguality by added translation tokens as new nodes and an additional translation edge type; responses from their system will only be from the nodes with translation edges. Their unsupervised approached is the most robust system in that they were the only system that attempted CLWSD for all five languages selected for the task in SemEval-2010. Surprisingly, their approach exceeded Mahapatra et al.'s (2010) knowledge-rich sense overlapping-similarity approach.

Topic Models for Word Sense Disambiguation

Topic models have been proposed by recent researchers for the WSD task. For example, Cai et al. (2007) exploited the corpus topics as global context features for WSD; Boyd-Graber et al. (2007) integrated McCarthy et al. (2004) approach for finding predominant word senses into a supervised topic modeling framework. Boyd-Graber and Blei also integrated their WSD system and it led to modest improvements to state-of-art information retrieval results. Li et al. (2010) proposed a probabilistic topic model based WSD method and they achieved state-of-art results on SemEval-2007's coarse-grain WSD evaluation.[3] Li et al.'s probabilistic topic model based WSD proposed a fully unsupervised WSD by reweighting the sense dominance with the product of the maximum conditional probability of a sense given the context and the introduction of the latent topic variable. However, all these WSD methods were evaluated on monolingual English WSD.

3.3 Thesis Approach to WSD

In this thesis, we explore the usage of topic models in crosslingual WSD task, we refer to the task as *Topical Crosslingual WSD*. In brief, the approach is to match query sentences to the training sentences and use the word-alignments as the response to the CLWSD query. The choice to use topic models is primarily based on its extensibility due to its probabilistic and knowledge-lean nature. Chapter 5 presents topic modeling in details and chapter 6 describes the Topical Crosslingual WSD task definition, implementation, results and conclusion.

[3] Li et al. (2010) @ 79.99% for all words; MFS @ 78.99, top performing supervised system (UoR-SSI) @ 83.21% (Navigli & Velardi, 2005)

Chapter 4

To disambiguate language meaning, human knowledge is emulated in the form of structured and unstructured machine-readable data called knowledge sources. These data are essential to associate conceptual senses to their word representation. They vary from corpora of texts with or without sense annotations to machine-readable dictionary, thesauri, ontologies, etc. This chapter provides a brief survey of notable knowledge resources of various degrees and concludes with the list of resources that the *Topical CLWSD* task uses

4.1 Unstructured resources

4.1.1 Corpora

Corpora are collections of texts used to model human language. Monolingual English corpora can be either (i) *raw*, i.e. unlabeled or (ii) *annotated* with Part-Of-Speech (POS) tagged, sense-annotated or other annotations useful in modeling language. Other than type (i) and (ii), parallel/comparable corpora have an additional alignment feature that is used to link sentences from the source language to the corresponding target language sentences.

Monolingual raw/pos-tagged corpora

- *Brown Corpus* (Francis & Kucera, 1979), a million word balanced corpus of American English texts, published in the 1961
- *British National Corpus* (Burnard, 2007), a 100 million word corpus of written and spoken (transcribed) samples of British English

- Wall Street Journal (WSJ) corpus (Charniak, 2000), a collection of 30 million words from the WSJ

- *WaCky (Web as Corpora Kool Yntiative) Corpora* are built by crawling websites from the .uk, .de and .it domains. The corpora contain more than a billion words each (Baroni et al., 2009).

- *Wikipedia dumps* as often used as raw corpora for WSD due to their size and availability (e.g. Li et al. (2011))

Monolingual sense-tagged corpora

- SemCor (Miller et al., 1994) is the most used sense-tagged corpus, it is a subset of the Brown Corpus (Francis & Kucera, 1964). It includes 352 texts tagged with 234,000 sense tags.

- *Defense Science Organization of Singapore (DSO) corpus* which includes 192,800 sense-tagged tokens of 191 words from the Brown and WSJ corpora (Ng & Lee, 1996b)

- *Open Mind Word Expert corpus* is made up of sense-tagged instances of 288 nouns collaboratively crowd-tagged by web users (Chklovski & Mihalcea, 2002).

- *Senseval* and *SemEval* data sets used in the various WSD evaluation tasks. For a historical and generic task overview of SemEval workshops, see `http://en.wikipedia.org/wiki/SemEval`. Also Navigli (2009) provided a detailed survey on the SemEval tasks and their respective competing systems.

Parallel raw/aligned corpora

- Europarl Corpus (Koehn, 2005) is a parallel and aligned corpus extracted from the proceedings of the European Parliament in 21 European languages. The size ranges from 1 million to 2.5 million sentences depending on the language pairs. The sentence alignments were processed using the Gale & Church (1993) algorithm.

- Tatoeba Corpus is a collaborative database of 2.2 million example sentences in 120 languages geared towards language learners. The example sentences are crowd-translated and moderated by web users (Breen, 2003).

Parallel sense-tagged corpora

- *MultiSemCor* is an English-Italian parallel corpus tagged with senses from the English and Italian wordnet (Bentivogli et al., 2004)

- *Japanese SemCor (JSemCor)* is a sense-tagged corpus of Japanese translated from the English SemCor texts; the Japanese senses are projected across from English WordNet. The corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged (Bond et al., 2012).

- *NTU-Multilingual Corpus* is a collection of parallel and aligned texts (made up of the *Cathedral and the Bazaar* corpus, *Dancing Man* corpus and texts from Singapore Tourism Board (STB) websites) tagged with English, Mandarin and Japanese WordNet senses. (Chapter 7 gives a brief introduction to the NTU-MC and its relation to this thesis).

4.1.2 Collocation Resources

Collocation resources (such as Word Sketch Engine,[1] JustTheWord[2] and Web1T 5-grams (Hawker et al., 2007)) registers sequence of words that co-occur more often than would be expected by chance. For example, *strong tea* (collocates) vs *powerful tea* (dispreferred).

4.1.3 Stoplists

Stoplists are lists of undiscriminating non-content words such as *a, an, the, he, she, etc.* Stopwords are removed in most IR/IE or NLP tasks for two main reasons; (i) the match between a query and a document should be based on felicitous terms (information) rather than high frequency non-content words (noise), (ii) the inverted file (i.e. the mapping from numbers to words) would be reduced by 30-50% (Manning et al., 2008).

[1] http://www.sketchengine.co.uk/

[2] http://www.just-the-word.com/

Dolamic & Savoy (2010) have shown that the usage of stoplist had significant improvement in search engine's document retrieval using the traditional Okapi BM25. Using other retrieval models, stoplists were able to reduce search while preserving mean average precision in document retrieval.

- *MySQL* query parser filters this list of English stopwords when users use full-text queries[3]

- *Lucene* siphons lists of stopwords for 33 different languages (ranging from Arabic, European, Scandinavian to Asian languages) when the *StopWordAnalyzerBase* is used in indexing or retrieving documents[4]

- *Snowball Tartarus* uses stoplists for 21 languages (Romance, Germanic and Scandinavia languages) for their text analyzers[5]

- *Rank.nl* is a popular Search Engine Optimization company that shares lists of stopwords for their article and page analyzers in 19 languages[6]

4.2 Structured resources

4.2.1 Machine-Readable Dictionaries and Thesauri

Machine-readable dictionaries (MRDs) were first made available in the 1980s and have since become a knowledge source for human-language modeling. Thesauri provide basic lexicographic relations, like synonymy, antonymy and possibly other semantic relations such as hypernymy, hyponymy, meronymy, etc. (Kilgarriff & Yallop, 2000)

- Longman Dictionary of Contemporary English (LDOCE) (Proctor, 1978) was the most widely used MRD for WSD before the widespread adoption of WordNet (Miller et al. (1990); Fellbaum (1998))

- *Roget's International Thesaurus* (Roget, 1852) classified words based on its relation to (i) abstract notions, (ii) space, (iii) matter, (iv) intellectual ideas, (v) volition and (vi) socio-emotional intuitions. The latest edition contains 250,000 entries.

[3] http://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html

[4] http://lucene.apache.org/core/4_0_0-BETA/analyzers-common/org/apache/lucene/analysis/util/StopwordAnalyze

[5] http://snowball.tartarus.org/

[6] http://www.ranks.nl/resources/stopwords.html

- *Macquarie Thesaurus* (Bernard, 1987) has more than 200,000 synonyms based on Australian English including Australian colloquialisms (e.g. *emo*) ,Multiword Expressions (e.g. *nutty as fruitcake*) and Aboriginal English (e.g. *booliman*).

4.2.2 Ontologies and Knowledge Bases

An ontology is a categorized set of concepts based on the relationships between concepts, it can be either domain-specific or generic. A Knowledge Base (KB) is an information repository that collects and organized referential knowledge about the world. KB extends ontologies' notion of categorizing concepts to include referential facts about a concept. Ontologies are useful in WSD to map related concepts. Graphically, entities (i.e. hypernyms of generic types like people, events, organization, etc.) are represented by nodes (e.g. `BobDylan` and `BlondeOnBlonde`), the edges are the relations (e.g. `created`) between two nodes, a fact consist of two nodes and their connecting edge (e.g. `BobDylan created BlondOnBlonde`)

Hypothetically, KBs can be used to constrict the search space of ultra-productive natural language into a finite set of conceptual units; thus improving robustness of WSD systems.

- *Suggested Upper Merged Ontology* (SUMO) is a generic ontology with meta-level concepts that do not belong to a specific domain. It contains 25,000 terms and 80,000 axioms (Niles & Pease, 2001b).

- *YAGO* is a semantic KB derived from Wikipedia, WordNet and SUMO contains more than 10 million entities and 120 million facts (Suchanek et al., 2007).

- *Freebase* is a user-generated KB consisting of structured data of well-known people, places and things. Each referential entity is a *topic* and they contain properties that has one or more values; and each topic-value pair constitute a fact in the KB. It can be visualized as a database made up of Wikipedia infoboxes like properties and values. Freebase contains over 38 million topics and 1.1 billion facts (Bollacker et al., 2008).

- *BabelNet* is a hybrid knowledge based built with YAGO and Freebase like name entities and also traditional hieratical ontology concepts.

(see chapter 3.1.1)

4.2.3 WordNets

A WordNet is a computational lexicon of psycholinguistic concepts and their respective word representations (*synsets*). Each concept identified by the lexicographer is tagged with a unique sense identification number (***ConceptID***).[7] (see section 2.2.1 for a brief history of WordNet). WordNets of different language exists in different sizes and licenses; the freer the license the more a wordnet is used (Bond & Paik, 2012).

- Princeton WordNet (Miller et al., 1990) is the original English WordNet, the latest version 3.0 contains 155,000 synsets covering 117,000 concepts.
- Open Multilingual WordNet (Bond & Pease, 2013) is a repository of WordNet with open source licenses from over 26 languages totaling to 100 thousand over concepts and 1.4 million word representations for these concepts.

4.3 Resources used for Topical CLWSD

The approach of Topical Crosslingual WSD is to use minimal knowledge to provide the best translation of an English word given a context sentence. Then we search the Open Multilingual Wordnet for the respective sense ID(s) of the polysemous English word and the translation word-pair.

The two main resources for the Topical Crosslingual WSD are (i) a parallel corpus and (ii) the Open Multilingual WordNet. The former is used to build topic models for the matching subtask and the latter use in the providing the concept ID for the mapping subtask (see chapter 5.1). Also, a minimalistic stoplists was used to reduce topic induction search space. The *rank.nl* (section 4.1.3) analyzers' stoplists because they were the shortest list as we want to preserve as many context words as possible for sense disambiguation.

[7] ConceptID exists mainly in 2 formats (i) Original Princeton format (e.g. ***dog%1:05:00::***, the pet) and (ii) Sense Offset, an 8 digit integer with POS tag (e.g. 02084071-n). For simplicity, we choose to use the latter for this thesis.

Chapter 5

Topic Models and Natural Language Processing

Human's perception and cognition can be understood through computational models that address particular human capacity such as memory, categorization, problem solving, pattern recognition, etc. (Marr (1982); Anderson & Bower (1974)). Human's inference and prediction have shown to reflect statistics of the environment; e.g. the probability that a memory will be needed correlates to the frequency of prior exposures (Anderson & Schooler, 1991). Sensitivity to relevant world statistics has been shown to guide cognitive judgment (e.g. Griffiths et al. (2006))

Hence, computationally and statistically we can emulate human's language ability using statistical inferred language model that assigns probability to sequence of words. The computational language modeling is achieved by:

(i) creating a probability distribution that can reflect humans' use of language,

(ii) calculating the probability of inferring certain psycholinguistically motivated language phenomena and

(iii) predicting language phenomena from the statistical inference when encountering new language data.

Iyer & Ostendorf (1999) posited that a topic can be represented as a probability distribution over words. For example, given the following sets of words (and its probability), we can infer that first set of words refer to the *biological plant* and the latter, *industrial plant.*

(0) $\{$*water*(0.038), *fertilizer*(0.027), *grow*(0.026), *tree*(0.024), *organic*(0.019), *industrial*(0.014), *pollution*(0.009), *strike*(0.003)$\}$

(1) $\{$*water*(0.038), *industrial*(0.027), *fertilizer*(0.026), *pollution*(0.024), *grow*(0.019), *strike*(0.014), *tree*(0.009),*organic*(0.003)$\}$

We refer to each set of words as a topic, z. And each document is a mixture of a fixed set of topics with varying probability, we refer to this mixture of topics per document as the gist, g. From a lexical semantic perspective we can perceive a topic as a cluster of related percepts (which may or may not pertain to any particular concept). And each document has varying possibilities of falling into different topics.

Hofmann (1999) pioneered the extraction of topics from large unannotated corpora using probabilistic latent semantic indexing (PLSI). Blei et al. (2003b) proposed generative topic modeling for collections of documents using Latent Dirichlet Allocation (LDA) by the introduction of a Dirichlet prior on the distribution over topics.

Other studies on using generative methods for topic induction includes Parametric Mixture Models (Ueda & Saito, 2006), variational Expectation Maximization (vEM) (Buntine, 2002); Chinese Restaurant Process (Blei et al., 2003a) Markov Chain Monte Carlo (MCMC) (Buntine & Jakulin, 2004), Gibb Sampling (Griffiths et al., 2007), etc.

The next section of this chapter describes the process in generating and inferring topic models using Latent Dirichlet Allocation (LDA).

5.1 Topic Models with Latent Dirichlet Allocation

Latent Dirichlet Allocation is the process of learning N topics from a text corpus and assigning the probability of every topics to each document[1] from the corpus.

The generative language models assume that in human language, latent semantic structures, l, generates words, w. Topic modeling posits that these latent semantic structures are in the form of *gists*. LDA's definition of topic model assumes that a document is generated by allocating a distribution over topics for each document's gist and then allocating the probability of each word given a topic determined by the gist (Blei et al., 2003b).

Formally, we define LDA as such. Given a multidocument corpus, of k size, expressed as vector of words $\mathbf{w} = \{w_1,...,w_n\}$, where w$_i$ belongs to d$_i$ in a word-document co-occurrence matrix, the *gist*, g, is multinomially distributed over T topics with $\theta^{(d)}$ (i.e. the probability of a particular topic occurring in document d for a given word) parameters, approximated to α under the Dirichlet distribution. The topics are represented by a multinomial distribution over words in the vocabulary, with $\phi_w^{(z)}$ (i.e. the probability of a particular word occurring in a particular topic), approximated to β under the Dirichlet distribution.

$$\mathrm{P}(z|g) = \theta^{(d)} \sim \text{Dirichlet}(\alpha)$$
$$\mathrm{P}(w|z) = \phi_w^{(z)} \sim \text{Dirichlet}(\beta)$$

In other words, the distribution of topics over words is assumed to be prior on a Dirichlet distributed alpha parameter. LDA users can specify the α and β hyperparameters that will affect the granularity of topics induced by the model (Griffiths & Steyvers, 2004).

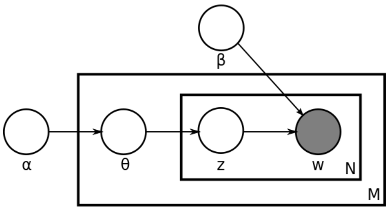


Figure 5.1: Graphical Representation of the Latent Dirichlet Allocation, where M = no. of documents and N = no. words per document

[1] Depending on the NLP task, a document can be as short as one sentence, a paragraph or as long as a full text

Below illustrate the intermediate outputs of the LDA topic modeling and Variational Bayesian topic inference. Firstly the a topic model is created by inducing a user-defined number of topics (e.g. no. of topics = 2).

Sample Corpus:

i. The trees grow with organic fertilizer and water
ii. The workers at the industrial plant that caused pollution are on strike
iii. That industrial plant manufactures fertilizer
iv. Remember to water the tree regularly

Each topic consist of feature words that are asummed to be describing the topic and each feature word is assigned a probability through the Latent Dirichlet Allocation process (Griffiths & Steyvers, 2004).

Induced Topics:

topic0: $\{$*water*(0.038), *fertilizer*(0.027), *grow*(0.026), *tree*(0.024), *organic*(0.019), *industrial*(0.014), *pollution*(0.009), *strike*(0.003)$\}$

topic1: $\{$*water*(0.038), *industrial*(0.027), *fertilizer*(0.026), *pollution*(0.024), *grow*(0.019), *strike*(0.014), *tree*(0.009),*organic*(0.003)$\}$

Finally, using the induced topics, the model infers the probability of a topic occuring for each document in the sample corpus. The probability is calculated using Variational Bayesian inference (Hoffman et al., 2010).

Document Gists:

i. (topic0 , 0.84) , (topic1, 0.16)
ii. (topic0 , 0.36) , (topic1, 0.64)
iii. (topic0 , 0.19) , (topic1, 0.81)
iv. (topic0 , 0.72) , (topic1, 0.28)

Chapter 6

Matching Query Sentence to Parallel Corpus using Topic Models for WSD

The `XLING` system introduces a novel approach to WSD by (i) first finding the closest match to the query sentences from sentences in a parallel corpus using topic models and it returns the word alignments as the translation for the target polysemous words, (ii) then using the word-translation pair we find the relevant sense ID to complete the sense disambiguation task. We call the first subtask (i) as the `match` and `translate` step and subtask (ii) as the `map` step.

This chapter examines the Topical CLWSD task by describing (i)`XLING` system's participation in the SemEval-2013 Crosslingual Word Sense Disambiguation (CLWSD) task (section 6.1 to 6.5), (ii) mapping the CLWSD outputs to WordNet senses and evaluating the overall accuracy of the Topical CLWSD approach (section 6.6). Thereafter this chapter concludes with discussion and the conclusion for the Topical CLWSD task. (section 6.7 - 6.8)

6.1 Background and Hypothesis

Topic modelling assumes that latent topics exist in texts and each semantic topic can be represented with a multinomial distribution of words and each document can be classified into different semantic topics (Hofmann, 1999). Blei et al. (2003b) introduced a Bayesian version of topic modeling using Dirichlet hyper-parameters, Latent Dirichlet Allocation (LDA). Using LDA, a set of topics can be generated to classify documents within a corpus. Each topic will contain a list of all the words in the vocabulary of the corpus with each word is assigned a probability of occurring given a particular topic (see chapter 2.2.3.2 and 3.2.3.2 for more detailed description of topic modeling and LDA)

We hypothesize that sentences with different senses of a polysemous word will be classified into different topics during the LDA process. By matching the query sentence to the training sentences by LDA induced topics, the most appropriate translation for the target polysemous word in the query sentence should be equivalent to translation of the word in the matched sentence(s) from a parallel corpus.

6.2 System Description

The `XLING_TnT` system attempts the matching subtask in three steps (1)`Topicalize`: matching the query sentence to the training sentences by the most probable topic. (2) `Rank`: the matching sentences were ranked according to the cosine similarity between the query and matching sentences. (3) `Translate`: provides the translation of the polysemous word in the matched sentence(s) from the parallel corpus.

6.2.1 Preprocessing

The Europarl corpus bitexts[1] (see chapter 4.1.1.3) were aligned at word-level with the `GIZA++` statistical word alignment tool. The translation tables from the word-alignments were used to provide the translation of the polysemous word in the Translate step.

[1] eng-deu, eng-spa, eng-fre, eng-ita, eng-nld

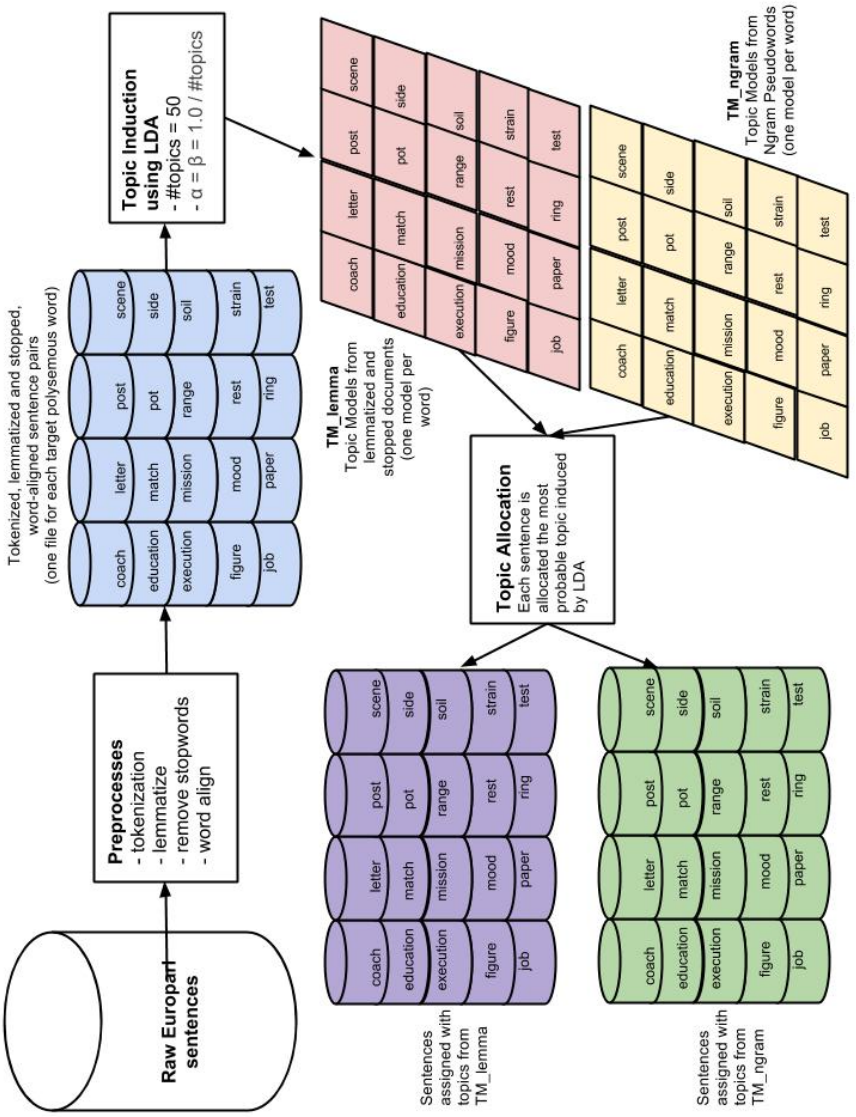

Raw Europarl sentences
Preprocesses
- tokenization
- lemmatize
- remove stopwords
- word align
Tokenized, lemmatized and stopped, word-aligned sentence pairs (one file for each target polysemous word)
coach
letter
post
scene
education
match
pot
side
execution
mission
range
soil
figure
mood
rest
strain
job
paper
ring
test
Topic Induction using LDA
- #topics = 50
- α = β = 1.0 / #topics
TM_lemma
Topic Models from lemmatized and stopped documents (one model per word)
TM_ngram
Topic Models from Ngram Pseudowords (one model per word)
Topic Allocation
Each sentence is allocated the most probable topic induced by LDA
Sentences assigned with topics from TM_lemma
Sentences assigned with topics from TM_ngram

Figure 6.1: XLING system flowchart: Training (section 6.2.1 to 6.2.2.2)

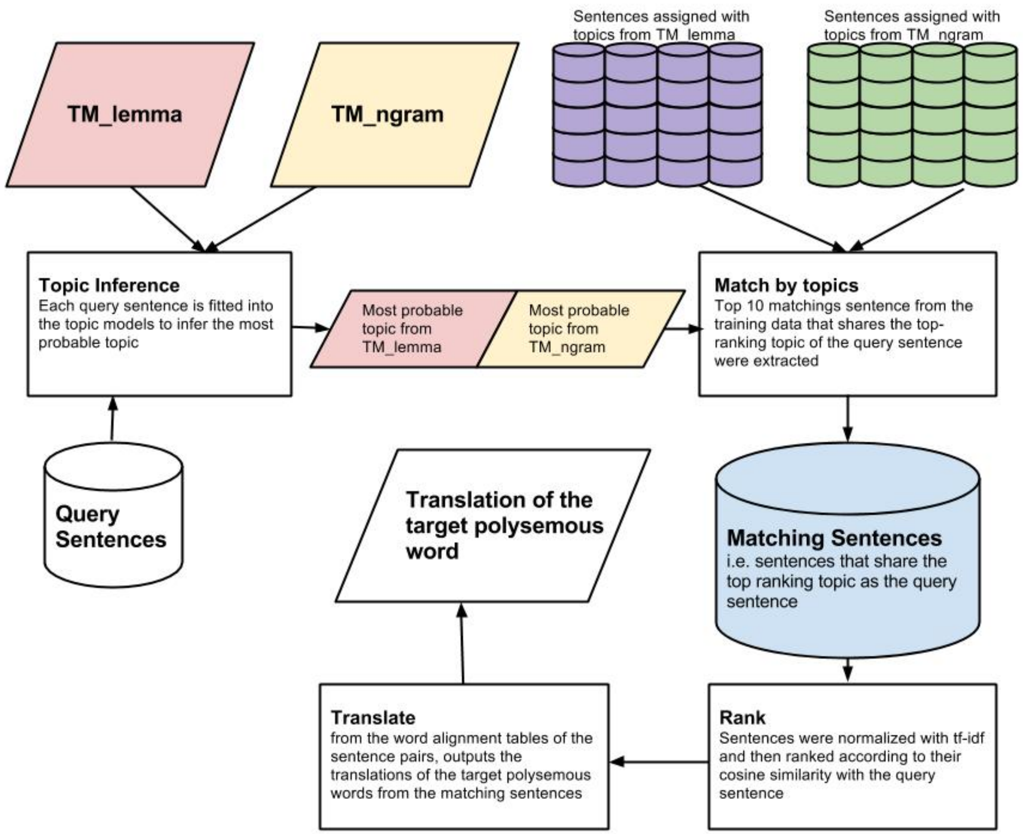

TM_lemma
TM_ngram
Sentences assigned with topics from TM_lemma
Sentences assigned with topics from TM_ngram
Topic Inference
Each query sentence is fitted into the topic models to infer the most probable topic
Most probable topic from TM_lemma
Most probable topic from TM_ngram
Match by topics
Top 10 matchings sentence from the training data that shares the top-ranking topic of the query sentence were extracted
Query Sentences
Translation of the target polysemous word
Matching Sentences
i.e. sentences that share the top ranking topic as the query sentence
Translate
from the word alignment tables of the sentence pairs, outputs the translations of the target polysemous words from the matching sentences
Rank
Sentences were normalized with tf-idf and then ranked according to their cosine similarity with the query sentence

Figure 6.2: XLING system flowchart: Testing (section 6.2.2.3 to 6.2.4)

The English sentences from the bitexts were lemmatized using a dictionary-based lemmatizer (`xlemma`).[2] The lemmatizer used WordNet entries as a lemma dictionary and it preferred tokens with plural if the plural and singular form of a word achieves two different synsets. For example, `xlemma` preferred spectacles (04272054-n) to spectacle (00075471-n, 06889138-n, 04271891-n). Before and after the lemmatization, English stopwords were removed from the sentences. The lemmatized and stop filtered sentences were used as document inputs to train the LDA topic model in the Topicalize step.

Previously, topic models had been incorporated as global context features into a modified naive Bayes network with traditional WSD features (Cai et al., 2007). We try a novel approach of integrating local context (N-grams) by using pseudo-word sentences as input for topic induction. For example:

Original Europarl sentence:

"If players fail to score, their coach does not go and widen the goal, but instead sees to it that they play better."

Lemmatized and stopped:

"player fail score coach go widen goal see play better

Ngram pseudo-word:

"player_fail_score fail_score_coach score_coach_go coach_go_widen go_widen_goal widen_goal_see goal_see_play see_play_better"

6.2.2 Topicalize

The `Topicalize` step of the system first (i) induces a list of topics using LDA, then (ii) allocates the topic with the highest probability to each training sentence and finally (iii) the topic is inferred from the query sentence and a list of sentences that shares the same top ranking topic are considered as matching sentences that contains the target polysemous word with the same sentence

6.2.2.1 Topic Induction

Topic models were trained using Europarl sentences that contain the singular or plural form of the focus word. The topic models were induced using

[2] https://code.google.com/p/xlemma/

LDA by setting the number of topics (#topics) as 50, and the alpha and beta hyper-parameters were symmetrically set at 1.0/#topics.[3] For example, *topic 1* is more relevant to the human coach (***09931640-n***) whereas *topic 2* is more relevant to the vehicular coach (***02924116-n***).

> *Topic 1*: [(0.0208, ’*sport*’), (0.0172, ’*player*’), (0.0170, ’*train*’),
> (0.0166, ’*league*’), (0.0133, ’*field*’), (0.0133, ’*football*’),
> (0.0130, ’bus’), (0.0117, ’*drive*’), (0.0117, ’*transport*’),
> (0.0130, ’tour’), (0.0117, ’*departs*’), (0.0117, ’*goal*’),
> (0.0111, ’*carriage*’)]

> *Topic 2*: [(0.0208, ’*bus*’), (0.0172, ’*drive*’), (0.0170, ’*train*’),
> (0.0166, ’*departs*’), (0.0133, ’*tour*’), (0.0133, ’*football*’),
> (0.0130, ’player’), (0.0117, ’*team*’), (0.0117, ’*transport*’),
> (0.0130, ’carriage’), (0.0117, ’*horse*’), (0.0117, ’*team*’),
> (0.0111, ’*sport*’)]

6.2.2.2 Topic Allocation

Each sentence was allocated the most probable topic induced by LDA. An induced topic contained a ranked list of tuples where the 2nd element in each tuple was a word that associated with the topic, the 1st element was the probability that the associated word will occur given the topic.

6.2.2.3 Topic Inference and Match

With the trained LDA model, each query sentence was fitted into the topic model to infer the probability of every induced topics given the query sentence. Using the most probable topic of each query sentence, sentences from the training corpus that shares the same top ranking topic was extracted. The top 10 training sentences that shared the top ranking topic were considered as matching sentences. The topic induction, allocation and inference were done separately on the lemmatized and pseudo-word sentence, resulting in two set of matching sentences. Only the sentences that were output from both set of matches are considered for the `Rank` step.

[3] Blei et al. (2003b) had shown that the perplexity plateaus at when #topics $\geq$ 50; higher perplexity meant more computing time needed to train the model.

6.2.3 Rank

A mini dictionary was built from the matching sentences output by Topicalize step and the documents were normalized using term frequency-inverse document frequency (tf-idf). Then the matching sentences were ranked according to the cosine similarity with the query sentences. Only the top five sentences were piped into the `Translate` step.

6.2.4 Translate

From the matching sentences, the `translate` step checks the GIZA++ (Och & Ney, 2003) word alignment table and outputs the translation(s) of the target polysemous word. Each matching sentence could output more than 1 translation depending on the target word alignment. As a simple way of filtering stop-words from target European languages, translations with less than 4 characters were removed. This effectively distilled misaligned non-content words, such as articles, pronouns, prepositions, etc. To simplify the lemmatization of Spanish and French plural noun suffixes, the -es and -s were stemmed from the translation outputs.

The `XLING_TnT` system output one translation for each query sentences for the best result evaluation and outputs the top 5 translations for the out-of-five result evaluation.

6.2.5 Fallback

For the out-of-five evaluation, if the query returned less than 5 answers, the first fallback appends the lemmas of the Most Frequent Sense (according to SemCor) of the target polysemous word in the respective language WordNets. The second fallback appended the most frequent translations of the target polysemous word to the queries response.

6.2.6 Baseline

Instead of matching sentences by topic models, we tried a simplistic baseline for matching sentences by cosine similarity between the lemmas sentence of the query sentence and the lemmas of each English sentence in the training corpus. The baseline system was named `XLING_SnT` (Similar and Translate). The cosine similarity was calculated from the division of the vector product of the query and training sentence (i.e. numerator) by the root product of the vectors magnitude squared.

6.3 CLWSD Task Evaluation

To empirically evaluate the matching subtask, we subscribed to the Crosslingual Word Sense Disambiguation (CLWSD) task from SemEval-2013 where 1000 queries from 20 polysemous:

- *coach*
- *education*
- *execution*
- *figure*
- *job*
- *letter*
- *match*
- *mission*
- *mood*
- *paper*
- *post*
- *pot*
- *range*
- *rest*
- *ring*
- *scene*
- *side*
- *soil*
- *strain*
- *test*

To score the system outputs, the CLWSD used the classic precision score[4] the mode precision from SemEval-2007's lexical substitution task (McCarthy et al., 2007)[5]. The mood precision was calculated based on the majority choice of translation that human annotators have selected. Let H be a set of human annotators, T be the set of query items with h_i as the set of responses for each query for each query $i \in T$ for annotator $h \in H$. For each $i \in T$, we calculated the mood (m_i) which corresponds to the translation with highest ranked translation from H. Let A be a set from T where the system provides at least one answer and $a_i : i \in A$ be the set of answers from the system for each query, i. For each i, we calculated the multiset union (H_i) for all h_i from all H. And for each unique answer (res) from H_i that has an associated frequency ($freq_{res}$). Three annotators were required to choose the top 3 translation of the polysemous target word for each query sentence, and the $freq_{res}$=1 if translation is picked by 1 annotator, $freq_{res}$=2 if by 2, and $freq_{res}$=3 if by 3 annotators.

The CLWSD task was evaluated on two sets of results *best* and *out-of-five* (*oof*). For the *best* evaluation, there was no limit on how many translation(s) the system can provide for the target polysemous word, but the score was divided by the number of answers. For the *oof* evaluation,

[4] precision = true positive / (true positive + false negative)

[5] The CLWSD refer to mode precision as mood precision or mood

systems can provide up to five answers and the score was not divided by the number of guesses.

$$P = \frac{\sum_{a_i : i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|A|}$$

$$Mood\ \ P = \frac{\sum_{a_i : i \in AM} 1\ if\ any\ guess \in a_i = m_i}{|AM|}$$

6.4 CLWSD Results

12 teams registered for the CLWSD task evaluation in SemEval-2013, 5 teams completed and submitted their evaluation results. Table 1 and 2 presents the results for the XLING system for best and out-of-five evaluation. Our system did worse than the tasks baseline, i.e. the Most Frequent Translation (MFT) of the target word for all languages. Moreover the topic model based matching did worse than the cosine similarity matching baseline. The results clearly disproved our hypothesis that sentences with different sentences with different senses of a polysemous word will be classified accordingly by topics during the LDA process.

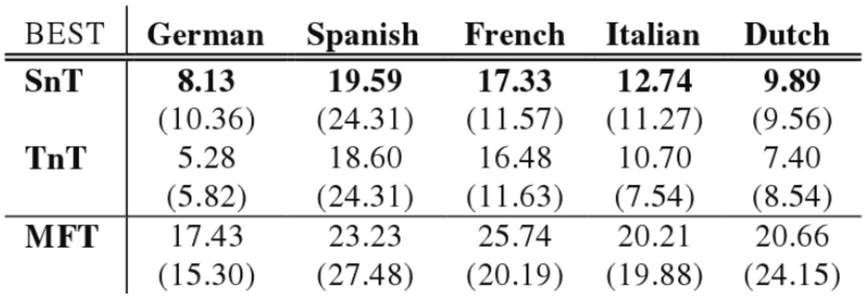
Li et al. (2010) and Anaya-Sánchez et al. (2007) had shown that pure topic model based unsupervised system for WSD should perform a little better than Most Frequent Sense baseline in coarse-grained English WSD. Hence it was necessary to perform error analysis and tweaking to improve the XLING system. Moreover, the XLING underperforms compared to other systems which completed the CLWSD task (see fig. 6.5).[6]

6.5 Error Analysis

Hyperparameters Modification

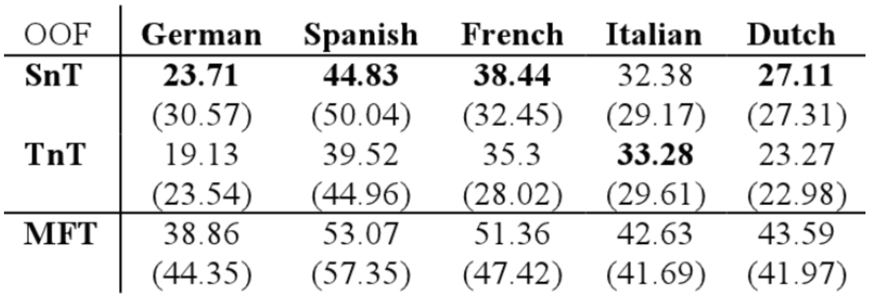
We also explored why the topic model based matching performed worse than surface cosine similarity matching. Statistically, we could improve the ro-

[6] ParaSense was the organizers' system



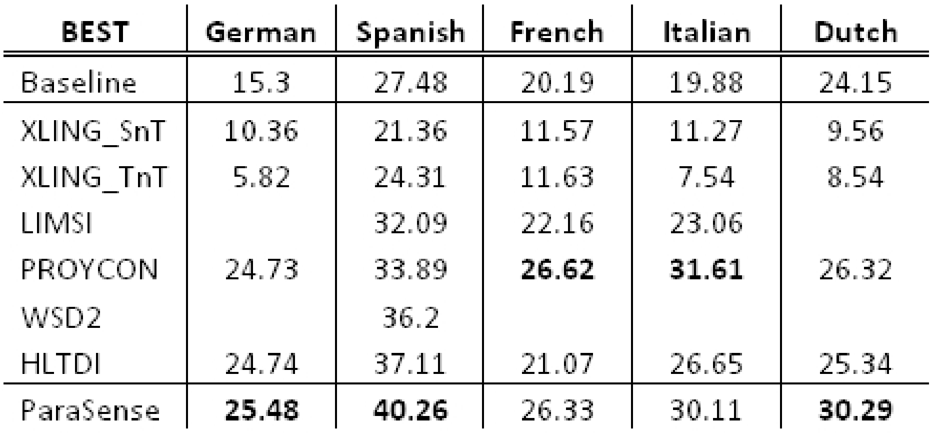
BEST	German	Spanish	French	Italian	Dutch
SnT	**8.13**	**19.59**	**17.33**	**12.74**	**9.89**
	(10.36)	(24.31)	(11.57)	(11.27)	(9.56)
TnT	5.28	18.60	16.48	10.70	7.40
	(5.82)	(24.31)	(11.63)	(7.54)	(8.54)
MFT	17.43	23.23	25.74	20.21	20.66
	(15.30)	(27.48)	(20.19)	(19.88)	(24.15)

Figure 6.3: Precision and (Mode) Precision for *best* Evaluation



OOF	German	Spanish	French	Italian	Dutch
SnT	**23.71**	**44.83**	**38.44**	32.38	**27.11**
	(30.57)	(50.04)	(32.45)	(29.17)	(27.31)
TnT	19.13	39.52	35.3	**33.28**	23.27
	(23.54)	(44.96)	(28.02)	(29.61)	(22.98)
MFT	38.86	53.07	51.36	42.63	43.59
	(44.35)	(57.35)	(47.42)	(41.69)	(41.97)

Figure 6.4: Precision and (Mode) Precision *oof* Evaluation

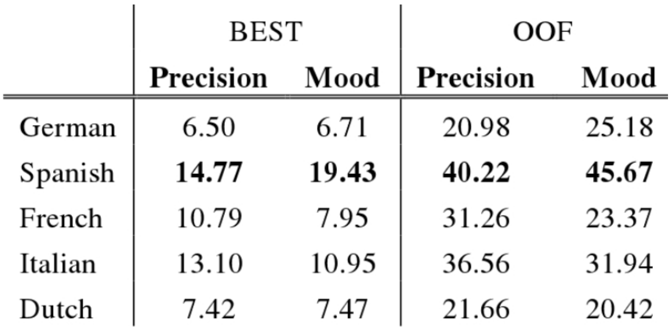


BEST	German	Spanish	French	Italian	Dutch
Baseline	15.3	27.48	20.19	19.88	24.15
XLING_SnT	10.36	21.36	11.57	11.27	9.56
XLING_TnT	5.82	24.31	11.63	7.54	8.54
LIMSI		32.09	22.16	23.06	
PROYCON	24.73	33.89	**26.62**	**31.61**	26.32
WSD2		36.2			
HLTDI	24.74	37.11	21.07	26.65	25.34
ParaSense	**25.48**	**40.26**	26.33	30.11	**30.29**

Figure 6.5: Mood Precision for *best* Evaluation of Competing Systems

bustness of the topic models by (i) tweaking the Dirichlet hyper-parameters to alpha = 50/#topics, beta = 0.01 as suggested by Wang et al. (2009).

Although the hyperparameter tweaks improved the scores for German and Dutch evaluations but the tweak brought the overall precision and mood precision of the other three languages down. Since the documents from each language were parallel, this poses the need to explore the level of language-dependency for LDAs hyperparameters.



	BEST		OOF	
	Precision	**Mood**	**Precision**	**Mood**
German	6.50	6.71	20.98	25.18
Spanish	**14.77**	**19.43**	**40.22**	**45.67**
French	10.79	7.95	31.26	23.37
Italian	13.10	10.95	36.56	31.94
Dutch	7.42	7.47	21.66	20.42

Figure 6.6: Evaluation Results of the TnT with Hyperparameter tweaks

Given no improvement from hyper-parameter tweaks, we reiterated Boyd-Graber et al.'s (2007) assertion that while topic models capture polysemous use of words, they do not carry explicit notion of senses that is necessary for WSD.

Data Sparsity and Redundency

Other than the hyperparameter issue, there were a couple of fail points that came from training data (a) data sparsity causing topic model fail and (b) topic drift that resulted in redundant overtraining (see Appendix A).

Across all languages, the queries for the polysemous word *pot* had retrieved poor precision (<5%), it is because the Europarl corpus had only 80-90 sentences with the word *pot*. One possible solution is to include sentences that contains synonyms of *pot* with only one sense.

However more data does not mean it is useful in training the topic models, e.g. for the Spanish evaluation of XLING_TnT (Figure A.2 from appendix), the model for the polysemous *job* (14 concepts) was trained from 10,000 documents while *test* (13 concepts) was trained with 2,800 documents but the XLING_TnT system scored 14.25% for *job* and 34.94% for *test*. We investigated the training document-precision correlation and when the topic model was overtrained, it caused instability in topic weighting and inference, it was not unlike the effects of topic drift in online LDA topic modeling (Hoffman et al., 2010).

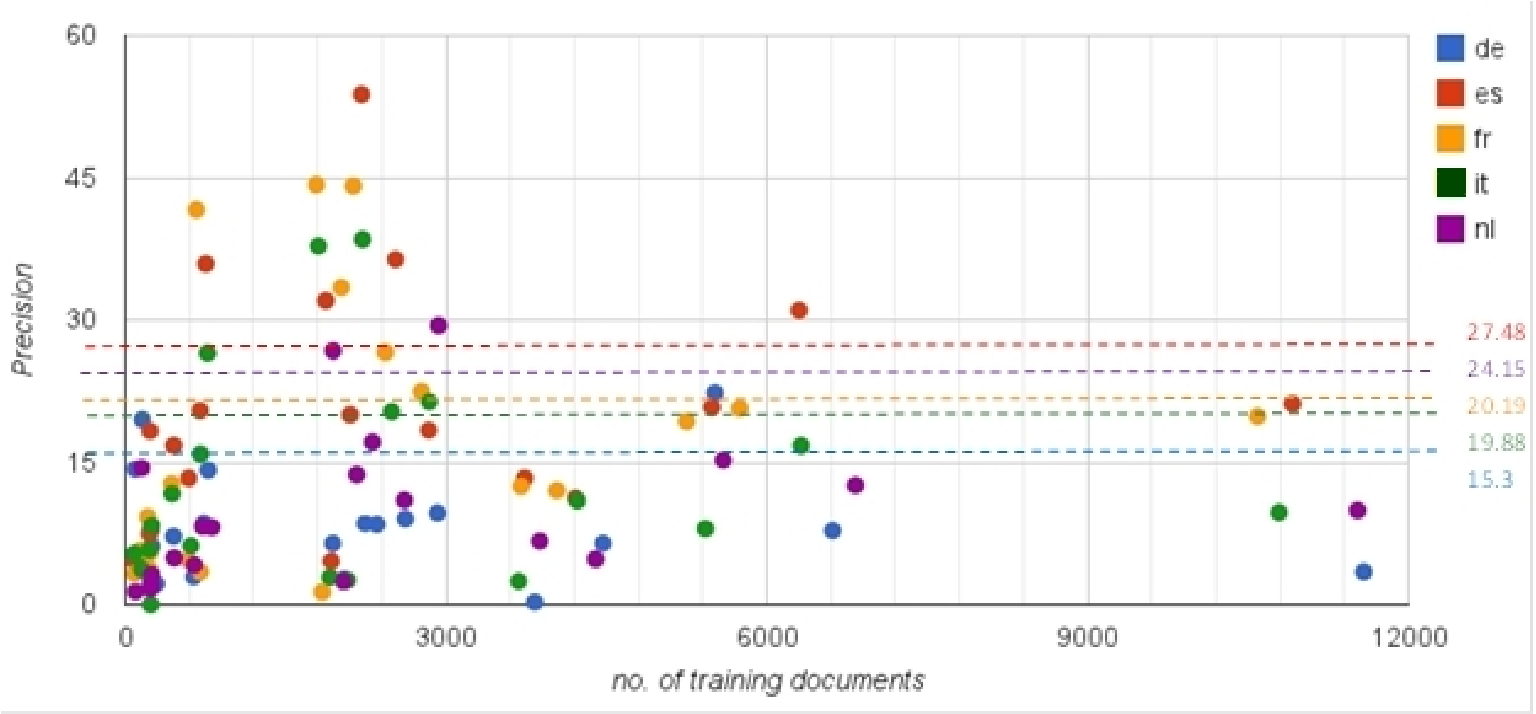


Figure 6.7: Overtraining: no. of Document-Precision Scatterplot

Figure 6.5 showed that precision falls below the baseline when model is trained with more than 3,000 training documents. This phenomenon was caused by our unconventional use of topic models; usually, topics were mined from large collection of documents and the topics were induced to represent the different topics that corpus represents, but we used LDA to induce finer grain topics that supposedly would fall into one single topic if mined from a large collection corpus. Thus, the hyperparameters, that determines granularity of topics, would differ and the optimal training corpus size would not be the presumed *as large as possible* size.

Not unlike various machine-learning statistical language modeling, the number of parameters, LDA has 50 free parameters[7] that the user can specify to train the topic model. Hence, finding the optimal values of each parameter requires much trial and error or mathematical estimation specific to the training data and task requirements.

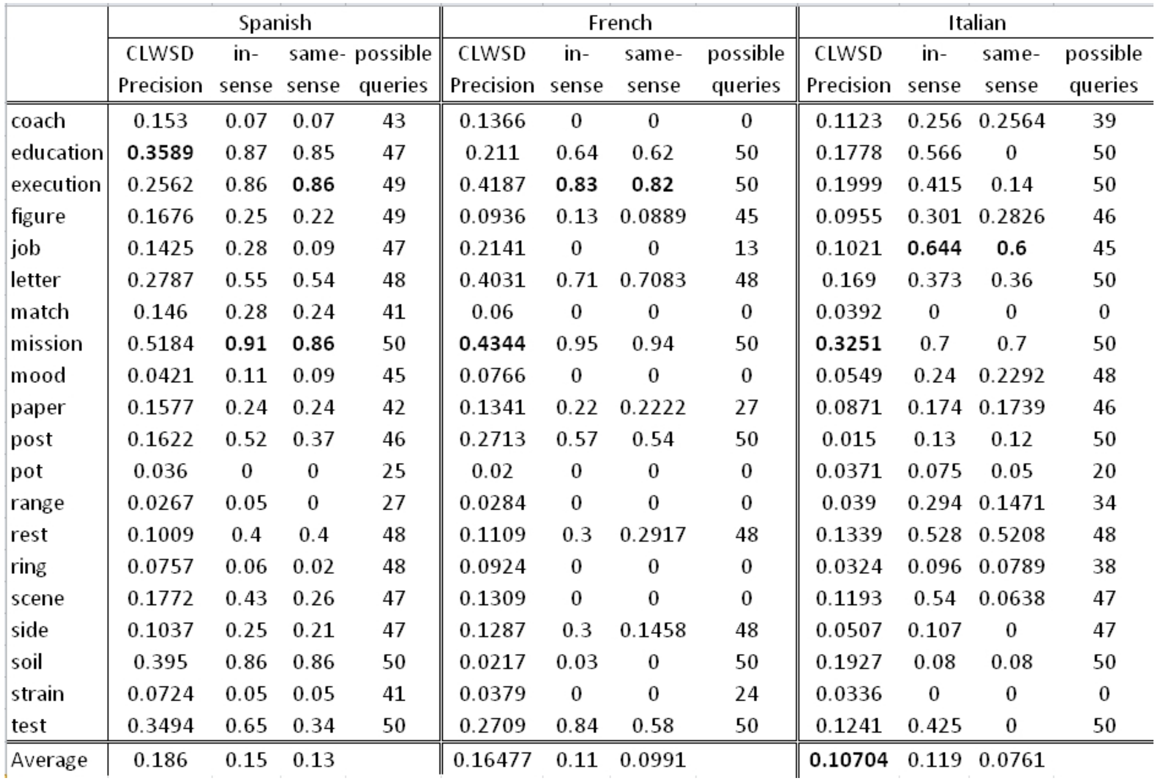
6.6 From CLWSD to Classic WSD

The second task of Topical CLWSD is to map the word-translation pair provided by the matching subtask to the Open Multilingual WordNet and return evaluate the answers based on whether their WordNet mapping are (i) a subset of the gold answer WordNet mapping (*insense*) or (ii) exactly the same as the gold answer mapping (*samesense*). This second task was only performed on the Romance languages (Spanish, French and Italian). For example, given the query sentence, Q:

> *"Then it's off aboard the **coach** into the reserve's safari park to see these animals, as well as others, roaming freely."*

The `XLING` system returns the best translation, *autobus*. And the Spanish gold answers for the query is (*autobus* and *autocar*). We map the gold answers to the OMW and produces *goldsense*, ***02924116-n*** and since the mapping from the system output only returns 1 sense, the *insame* and *samesense* score is both 100%. The *insame* and *samesense* score average per polysemous word per language is presented on figure 6.6.

[7] To name a few, #topics, α, β, random seed value to initialize the generative process (i.e. Expectation Maximization for LDA), etc.



	Spanish				French				Italian			
	CLWSD Precision	in-sense	same-sense	possible queries	CLWSD Precision	in-sense	same-sense	possible queries	CLWSD Precision	in-sense	same-sense	possible queries
coach	0.153	0.07	0.07	43	0.1366	0	0	0	0.1123	0.256	0.2564	39
education	**0.3589**	0.87	0.85	47	0.211	0.64	0.62	50	0.1778	0.566	0	50
execution	0.2562	0.86	**0.86**	49	0.4187	**0.83**	**0.82**	50	0.1999	0.415	0.14	50
figure	0.1676	0.25	0.22	49	0.0936	0.13	0.0889	45	0.0955	0.301	0.2826	46
job	0.1425	0.28	0.09	47	0.2141	0	0	13	0.1021	**0.644**	**0.6**	45
letter	0.2787	0.55	0.54	48	0.4031	0.71	0.7083	48	0.169	0.373	0.36	50
match	0.146	0.28	0.24	41	0.06	0	0	0	0.0392	0	0	0
mission	0.5184	**0.91**	**0.86**	50	**0.4344**	0.95	0.94	50	**0.3251**	0.7	0.7	50
mood	0.0421	0.11	0.09	45	0.0766	0	0	0	0.0549	0.24	0.2292	48
paper	0.1577	0.24	0.24	42	0.1341	0.22	0.2222	27	0.0871	0.174	0.1739	46
post	0.1622	0.52	0.37	46	0.2713	0.57	0.54	50	0.015	0.13	0.12	50
pot	0.036	0	0	25	0.02	0	0	0	0.0371	0.075	0.05	20
range	0.0267	0.05	0	27	0.0284	0	0	0	0.039	0.294	0.1471	34
rest	0.1009	0.4	0.4	48	0.1109	0.3	0.2917	48	0.1339	0.528	0.5208	48
ring	0.0757	0.06	0.02	48	0.0924	0	0	0	0.0324	0.096	0.0789	38
scene	0.1772	0.43	0.26	47	0.1309	0	0	0	0.1193	0.54	0.0638	47
side	0.1037	0.25	0.21	47	0.1287	0.3	0.1458	48	0.0507	0.107	0	47
soil	0.395	0.86	0.86	50	0.0217	0.03	0	50	0.1927	0.08	0.08	50
strain	0.0724	0.05	0.05	41	0.0379	0	0	24	0.0336	0	0	0
test	0.3494	0.65	0.34	50	0.2709	0.84	0.58	50	0.1241	0.425	0	50
Average	0.186	0.15	0.13		0.16477	0.11	0.0991		**0.10704**	0.119	0.0761	

Figure 6.8: CLWSD to Classic WSD

Another example with query sentence:

"While the ***test*** *results for physical search exceeded the national average, both the metal detector and X-ray results were below average."*

The system answer is *prueba* which returns three possible senses {00791078-n,05799212-n,00794367-n}. And the gold answers (ensayo;experimento; inspeccion;investigacion;prueba;test) returns 5 possible senses {07197021-n, 01006675-n, 00791078-n, 05799212-n, 00794367-n}. The *insense* score would be 0.6 while the *samesense* score would be 0.

Figure 6.6 showed getting the same sense remains a challenge due to the low precision of the matching task. The *insense* and *samesense* scores for French and Italian were lower than the Spanish because the possibility of mapping the gold answers provided by the CLWSD to the OMW is lower (see *possible query* in figure 6.6). Given sufficient coverage of the WordNet, it might be possible to achieve sense mappings that produce samesense scores that are close to the CLWSD precision of the system.

By disregarding the unmapped senses, we partially overcame the lack of coverage of the target language WordNets but failed to capture correct senses answered by the system that were not in the OMW. The previous insense and samesense score can be comparable to performance of systems participating in the CLWSD task. To make the traditional precision, recall and f-score measure comparable to classic WSD task[8], we excluded the queries that did not return sense mapping from the OMW and take the *samesense* counts as true positive (*tp*), wrong answers as the false positive (*fp*). If the system outputs a senseID although the gold translation could not, it also counted towards false negative (*fn*) score.

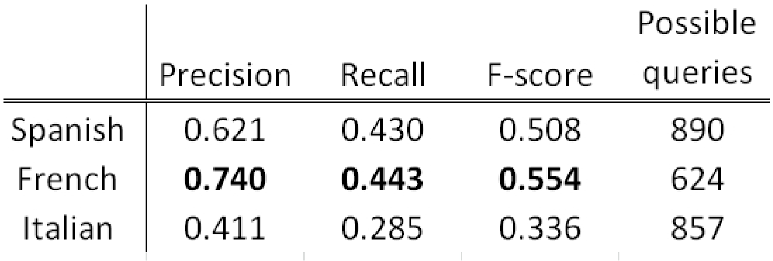
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Figure 6.7 presented the precision, recall and f-score on the adapted classic WSD task using gold translation mappings from the OMW as the gold

[8] more specifically, classic coarse-grain WSD, since the mapping a gold tranlsation from CLWSD to OWN produces more than one senseID



	Precision	Recall	F-score	Possible queries
Spanish	0.621	0.430	0.508	890
French	**0.740**	**0.443**	**0.554**	624
Italian	0.411	0.285	0.336	857

Figure 6.9: Topical CLWSD score (comparable to Classic WSD)

senses and the OMW mappings of the XLING_TnT outputs from the previous CLWSD task. This OMW mapped WSD subtask is very much dependent on the performance of the system in the CLWSD task as well as the coverage of the various target language WordNets. For the French queries, we achieved high precision because the number of mappings from the gold translations of the CLWSD to the OMW is almost halved the number of all the possible queries.[9]. For the Italian queries, we scored low on both precision and recall mainly due to the lack of precision in the CLWSD task.

The results showed that by crosslingual approach to WSD was able to achieve high precision score and reinstated the usage of parallel texts to reduce the sense ambiguity of polysemous words. However the recall remains low due to number of false negatives, mainly partial answers from *insense* that XLING system provides; i.e. mainly because the outputs from the CLWSD substask could not be mapped to the Open Multilingual WordNet.

6.7 Discussion

The main advantage of statistical language independent approaches is the ability to scale the system in any possible language. However language dependent processing remains crucial in providing accuracy to the system, especially lemmatization in WSD tasks (e.g. *kraftomnibusverkehr*). Moreover, disambiguating senses solely from sentential context is artificially hard. By going through the individual queries and responses from the matching substask, we identified several issues in the translate step that needed to be resolved to achieve higher precision. Using language specific lemmatiz-

[9] The maximum number of possible queries is 1000, all 50 queries for each of the 20 polysemous words in the CLWSD task

ers, as other competing systems for the CLWSD task had done, would had improved the accuracy of the system. For example:

(i) German-English and Dutch-English word alignments containing compound words needed to be segmented (e.g. *kraftomnibusverkehr* needed to be segmented to *kraft omnibus verkehr* and realigned such that the target word *coach* only aligns to *omnibus*),

(ii) de-pluralization of Italian, German and Dutch was crucial was getting the gold answers of the task (e.g. `XLING` answers *omnibussen* while the gold answers allowed *omnibus*).

6.8 Conclusion for Topical CLWSD

This thesis had defined the Topical CLWSD task by first finding a match to the query sentence from a parallel corpus and disambiguating the senses from the interlinked Open Multilingual Wordnet (OMW). The topic-model based sentence matching fails to meet the Most Frequent Translation (MFT) baselines. But the surface cosine baseline, without any incorporation of any sense knowledge, had surprisingly achieved performance closer to MFT. The surface cosine similarity measure could serve as a baseline for the sentence matching subtask in future research.

Although the sentence matching subtask required much attention for future research, the disambiguation of senses from OMW has shown potential in WSD if query sentences from the WSD evaluation tasks contain not just the English sentences but also their translations in multiple languages.

Chapter 7

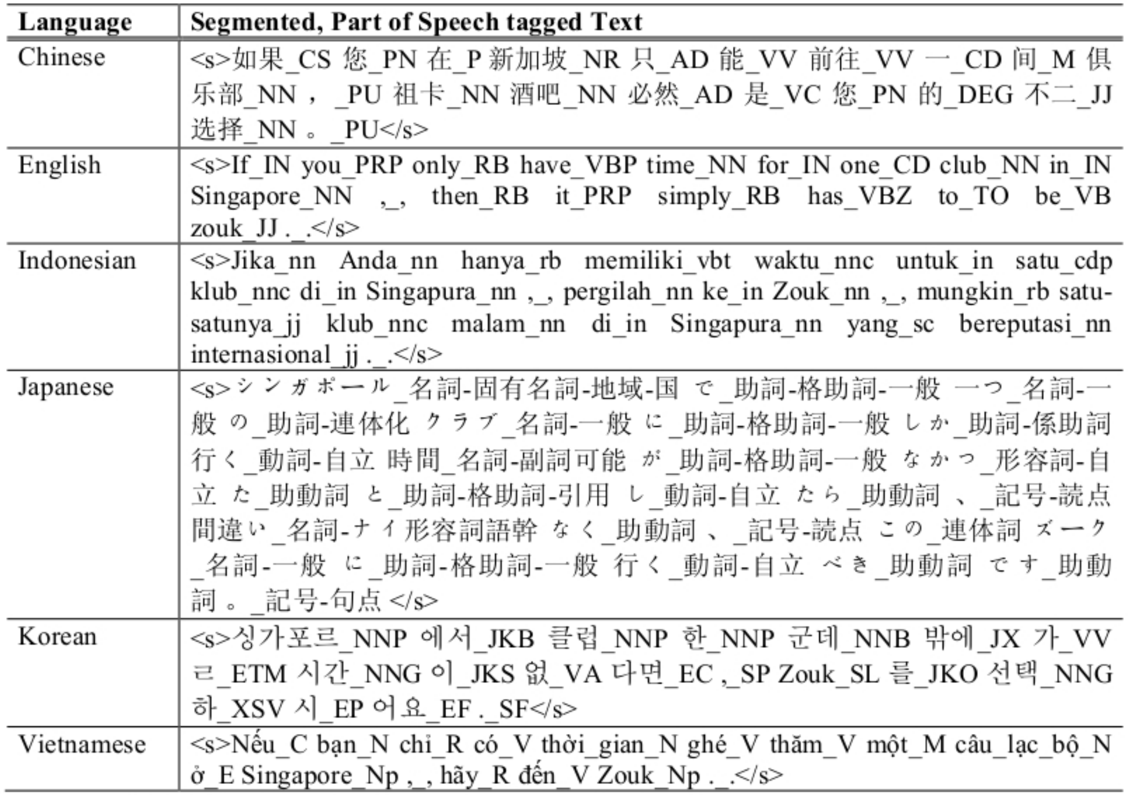
Nanyang Technological University - Multilingual Corpus (NTU-MC)[1]

WSD is not a stand-alone task and neither is the XLING system. The XLING system serves as one of the language disambiguation devices in the quest to understand human language (syntactically and semantically) crosslingually. The pursuit of crosslingual language technologies is based on the realization that parallel/translated/comparable texts from linguistically diverse languages provide more information in human language understanding for NLP tasks. Thus, crosslingual NLP researches focus on (i) improving multilingual data size and annotations and (ii) extracting disambiguating knowledge from multilingual data.

7.1 Linguistically Diverse Corpus

The NTU-MC is a linguistically diverse corpus that contains 595,000 words (26,000 sentences) in 7 languages (Arabic, Chinese, English, Indonesian, Japanese, Korean and Vietnamese) from 7 different language families (Afro-Asiatic, Sino-Tibetan, Indo-European, Austronesian, Japonic, Korean (a language isolate) and Austro-Asiatic). The current version of the NTU-MC consists of text from two subcorpora (`yoursingapore` and `singaporemedicine`).

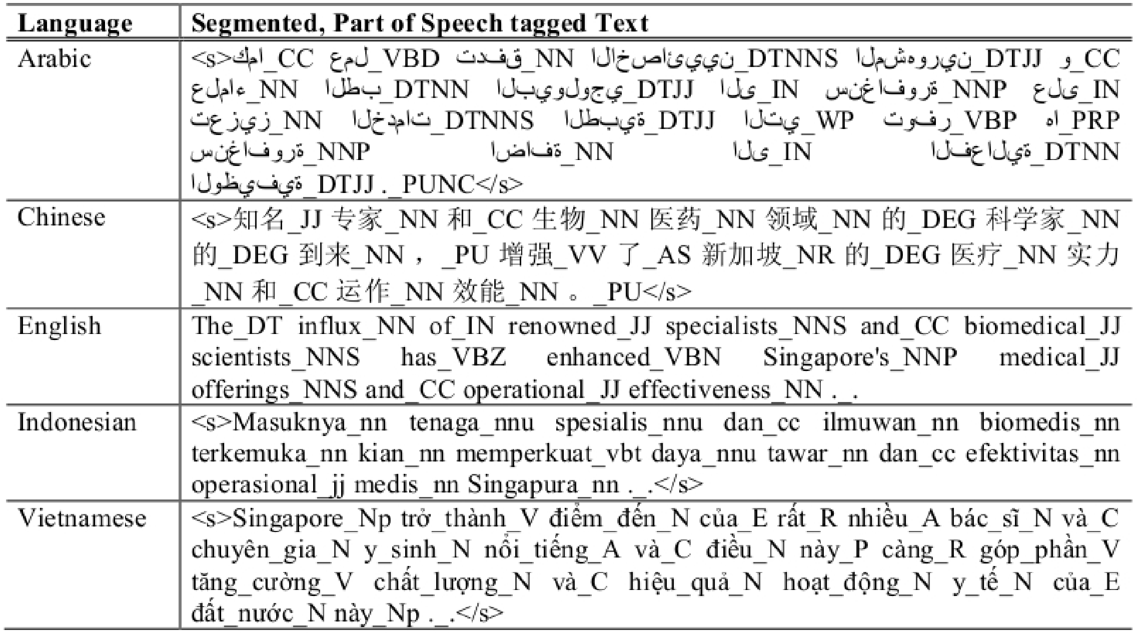
[1] Earlier versions of this chapter appear in Tan & Bond (2011, 2012)



Language	Segmented, Part of Speech tagged Text
Chinese	<s>如果_CS 您_PN 在_P 新加坡_NR 只_AD 能_VV 前往_VV 一_CD 间_M 俱乐部_NN ， _PU 祖卡_NN 酒吧_NN 必然_AD 是_VC 您_PN 的_DEG 不二_JJ 选择_NN 。_PU</s>
English	<s>If_IN you_PRP only_RB have_VBP time_NN for_IN one_CD club_NN in_IN Singapore_NN ,_, then_RB it_PRP simply_RB has_VBZ to_TO be_VB zouk_JJ ._.</s>
Indonesian	<s>Jika_nn Anda_nn hanya_rb memiliki_vbt waktu_nnc untuk_in satu_cdp klub_nnc di_in Singapura_nn ,_, pergilah_nn ke_in Zouk_nn ,_, mungkin_rb satu-satunya_jj klub_nnc malam_nn di_in Singapura_nn yang_sc bereputasi_nn internasional_jj ._.</s>
Japanese	<s>シンガポール_名詞-固有名詞-地域-国 で_助詞-格助詞-一般 一つ_名詞-一般 の_助詞-連体化 クラブ_名詞-一般 に_助詞-格助詞-一般 しか_助詞-係助詞 行く_動詞-自立 時間_名詞-副詞可能 が_助詞-格助詞-一般 なかっ_形容詞-自立 た_助動詞 と_助詞-格助詞-引用 し_動詞-自立 たら_助動詞 、_記号-読点 間違い_名詞-ナイ形容詞語幹 なく_助動詞 、_記号-読点 この_連体詞 ズーク_名詞-一般 に_助詞-格助詞-一般 行く_動詞-自立 べき_助動詞 です_助動詞 。_記号-句点 </s>
Korean	<s>싱가포르_NNP 에서_JKB 클럽_NNP 한_NNP 군데_NNB 밖에_JX 가_VV ㄹ_ETM 시간_NNG 이_JKS 없_VA 다면_EC ,_SP Zouk_SL 를_JKO 선택_NNG 하_XSV 시_EP 어요_EF ._SF</s>
Vietnamese	<s>Nếu_C bạn_N chỉ_R có_V thời_gian_N ghé_V thăm_V một_M câu_lạc_bộ_N ở_E Singapore_Np ,_, hãy_R đến_V Zouk_Np ._.</s>

Table 2: A sample of monolingual annotation from yoursingapore

Figure 7.1: A sample of monolingual annotation from yoursingapore



Language	Segmented, Part of Speech tagged Text
Arabic	<s>كما_CC عمل_VBD تدفق_NN الاختصاصيين_DTNNS المشهورين_DTJJ و_CC علماء_NN الطب_DTNN البيولوجي_DTJJ الى_IN سنغافورة_NNP على_IN تعزيز_NN الخدمات_DTNNS الطبية_DTJJ التي_WP توفر_VBP ها_PRP سنغافورة_NNP اضافة_NN الى_IN الفعالية_DTNN الوظيفية_DTJJ ._PUNC</s>
Chinese	<s>知名_JJ 专家_NN 和_CC 生物_NN 医药_NN 领域_NN 的_DEG 科学家_NN 的_DEG 到来_NN ，_PU 增强_VV 了_AS 新加坡_NR 的_DEG 医疗_NN 实力_NN 和_CC 运作_NN 效能_NN 。_PU</s>
English	The_DT influx_NN of_IN renowned_JJ specialists_NNS and_CC biomedical_JJ scientists_NNS has_VBZ enhanced_VBN Singapore's_NNP medical_JJ offerings_NNS and_CC operational_JJ effectiveness_NN ._.
Indonesian	<s>Masuknya_nn tenaga_nnu spesialis_nnu dan_cc ilmuwan_nn biomedis_nn terkemuka_nn kian_nn memperkuat_vbt daya_nnu tawar_nn dan_cc efektivitas_nn operasional_jj medis_nn Singapura_nn ._.</s>
Vietnamese	<s>Singapore_Np trở_thành_V điểm_đến_N của_E rất_R nhiều_A bác_sĩ_N và_C chuyên_gia_N y_sinh_N nổi_tiếng_A và_C điều_N này_P càng_R góp_phần_V tăng_cường_V chất_lượng_N và_C hiệu_quả_N hoạt_động_N y_tế_N của_E đất_nước_N này_Np ._.</s>

Table 3: A sample of monolingual annotation from singaporemedicine

Figure 7.2: A sample of monolingual annotation from singaporemedicine

The NTU-MC is annotated with a layer of monolingual annotation (POS and sense tags) and crosslingual annotation (sentence level alignments). The diverse language data and crosslingual annotations provide valuable information on linguistic diversity for traditional corpus linguistic research as well as natural language processing tasks such as Machine Translation. The NTU-MC is an on-going effort to compile and redistribute machine-readable and quality annotations from parallel/translated texts with linguistically diverse languages.

7.1.1 Historical log

The first version of the NTU-MC consisted of the foundation texts from Singapore Tourism Board's (STB) website (`www.yoursingapore.com`). It was built in May 2011 with 375,000 words (15,000 sentences) in 6 languages from 6 different language family trees. Part-of-Speech annotations were included in various degrees of accuracy. Other than being machine-readable, the corpus was designed to suit the Corpus Query Processor web (CQPWeb) graphic user interface.

The second version of the NTU-MC (Dec, 2011) moves on from monolingual POS annotations to provide crosslingual sentence alignments useful for multilingual NLP tasks such as machine-translation or language detection. English-Chinese and English-Japanese bitexts were manually aligned at sentence level and the other language pairs were automatically using HunAlign (Varga et al. 2005).

The third version of the NTU-MC (May, 2012) had an increased size with texts from STB's medical tourism website (`www.singaporemedicine.com`). With the inclusion of the new subcopus, the NTU-MC grew to 595,000 words (26,000 sentences) with the addition of Arabic texts from the new subcorpus. We have also implemented a probabilistic Bahasa Indonesian POS-tagger with the specifications recommended by Pisceldo et al. (2010).

The forth version of the NTU-MC (Jan, 2013) focused on improving annotations' accuracy. The yoursingapore subcorpus was manually sense-tagged (40,000 concepts for each language) in English, Chinese, Indonesian and Japanese. To resolve mis-segmentation issues as reported in the previous versions of NTU-MC, a dictionary of Singaporean street names, train-stations and landmarks were created with a dictionary-based Chinese segmentation tool (`minisegmenter`).

The fifth version of the NTU-MC is scheduled to be release in September 2013 with automatically sense-tagged annotations with the XLING system, as well as the merger of two new subcorpora ('dancing man' and 'cathedral and bazaar').

7.2 Asian Language NLP tools

Along with the corpus compilation, NLP tools were created to annotate the corpus. Although NLP tools for Asian Languages are available, they are seldom accessible nor given attention and often lack documentation.

7.2.1 Indonesian POS tagger

IndoTag is a probabilistic Conditional Random Field (CRF) Bahasa Indonesian Part of Speech (POS) tagger with the specifications recommended by Pisceldo et al. (2010). The pre-trained model is based on the unigram CRF with 2-left and 2-right context features using the Universitas Indonesia's 1 million word corpus compiled under the Pan Asia Networking Localization (PANL10N) project.[2] The IndoTag achieved 78% accuracy in a text sample POS evaluation of the NTU-MC.

7.2.2 Mini-segmenter

Mini-segmenter[3] (`mini-segmenter`) is a lightweight lexicon/dictionary based Chinese text segmenter; it adds whitespace to separate and tokenize the text. The advantage of using a lexicon/dictionary for text segmentation is the ability to localize and scale according to the text's language or domain. Supporting the open source movement, the default dictionary used by mini-segmenter is MDBG's (2013) CC-CEDICT.

The `mini-segmenter` first generates all possible combinations of tokens using the dictionary and ranks the combinations according to a ad-hoc scoring system (i.e. `mini-square` score). It is calculated using the summation of the square of the length of each segment. This novel scoring is based on the preference for larger chunks than smaller chunks in a sentence.

[2] PANL10N project is an initiative to build capacity in regional institutions for local language computing in South and South-East Asian countries

[3] The `mini-segmenter` at https://code.google.com/p/mini-segmenter/

7.2.3 GaChalign

GaChalign[4] is a python implementation of Gale-Church's (1993) length-based sentence aligner with options for variable parameters (viz. mean, variance, penalty). The aim of the tool development was to address the issue of poor sentence alignment between logographic/syllabic languages to alphabetic languages. Our experiment with English-Japanese NTU-MC bitext has shown that:

- aligning a syllabic/logographic language (JPN) to an alphabetic language (ENG) remains a challenge for Gale-Church algorithm (f-score peaks at 62.9%)
- using the calculated character mean[5] from the unaligned text improves precision and recall of the algorithm
- using the calculated alignment type penalties from a sample gold corpus also improves fscore

[4] GaChalign is freely available at https://code.google.com/p/gachalign/

[5] i.e. the ratio of source language characters to target language characters

Chapter 8

Conclusion and Future Work

Language is ambiguous and understanding language ambiguity computationally remains a challenge for the current state-of-art technologies. In this thesis we attempted to resolve language ambiguity through the Word Sense Disambiguation (WSD) task using parallel text and statistical topic models. Different from previous approaches to WSD, we explore the disambiguating power of using parallel text in the WSD task by (i) first finding a match to an English sentence with a polysemous word from a parallel corpus and extract the tranlsation of the polysemous word from the corpus' word-alignment (`match` and `translate`) then (ii) mapping the word-translation pair to the Open Multilingual WordNet (OMW) (`map`) and evaluate the sense disambiguation system base on the OMW sense inventory. We named the task topical CLWSD.

The `XLING` system was built to attempt the topical CLWSD task. Although the `XLING` system performed under the Most Frequent Translation (MFT) baseline in the first sentence-matching subtask evaluation[1], but it was able to achieve high precision on the OMW-mapped WSD subtask.

Future research on WSD and meaing disambiguation should further exploit the use of parallel translations to reduce the sense disambiguity as we have shown in the OMW-mapped WSD subtask. Crosslingual knowledge become increasingly important in the field of machine translation where crosslingual semantic/syntactic information has proven to improve performance.

[1] We adopt the SemEval-2013's Crosslingual WSD (CLWSD) evaluation and to evaluate the sentence-matching subtask

Other than computationally disambiguating word senses, this thesis updated the readers on the ongoing work in compiling the linguistically diverse Nanyang Technological University - Multilingual Corpus (NTU-MC). Both the `XLING` and the NTU-MC are related in their attempts to build crosslingual technologies.

Future Work

It is important to note that the thesis' approach to WSD involved minimal usage of knowledge resources and preprocessing tools; the XLING system used an aligned parallel corpus, an English lemmatizer and two sentence similarity measures (topic based matching and cosine similarity) and the Open Multilingual WordNet. Topical CLWSD's (i) `match` and `translate` and (ii) `map` approach to WSD shows potential when we consider the high precision that can be yielded by the `map` step (i.e. generating the overlapping synsets of the polysemous words and their matched translations).

Future work on the `match`, `translate` and `map` approach should attend to the low precision and mood of the `match` step, where the CLWSD system finds a match to the context sentence from a parallel corpus.

In the Topical CLWSD experiment, sentences were allocated only the top ranking LDA induced topic. In the light of the results, using the full topic distribution to compute similarities between queries and the training corpus might provide a richer representation to better capture sense distinction and improve the results.

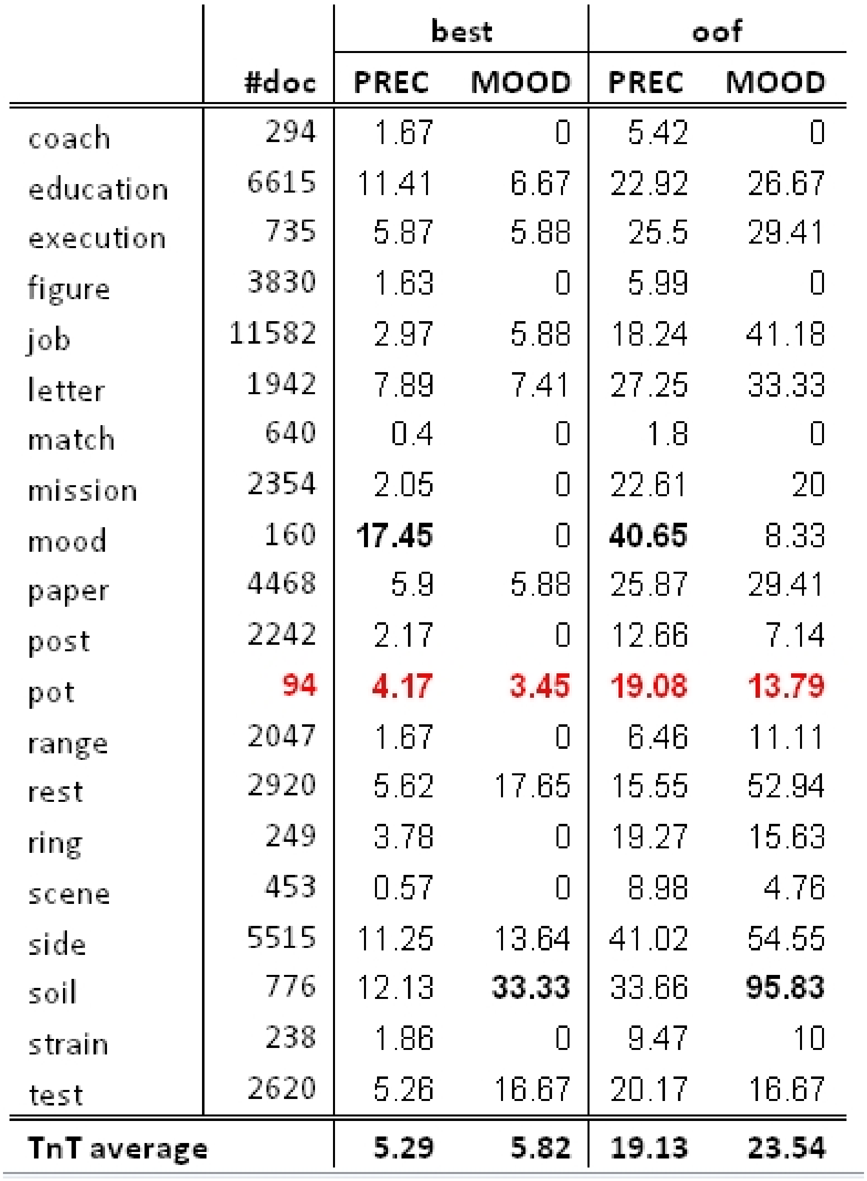
The XLING system used sentential cosine similarity measure to find the sentences from the training corpus that are semantically similar to the context sentence. It may be beneficial to compare various groupwise semantic similarity measure (e.g. Jaccard index, Normalized Google distance) to determine which measure is most apt for the CLWSD task.

Other than improving the similarity matching measure, it might be simpler to consider using pre-existing paraphrasing software (e.g. SEMILAR) to assess the similarity between the context sentence and the corpus.

Appendix A

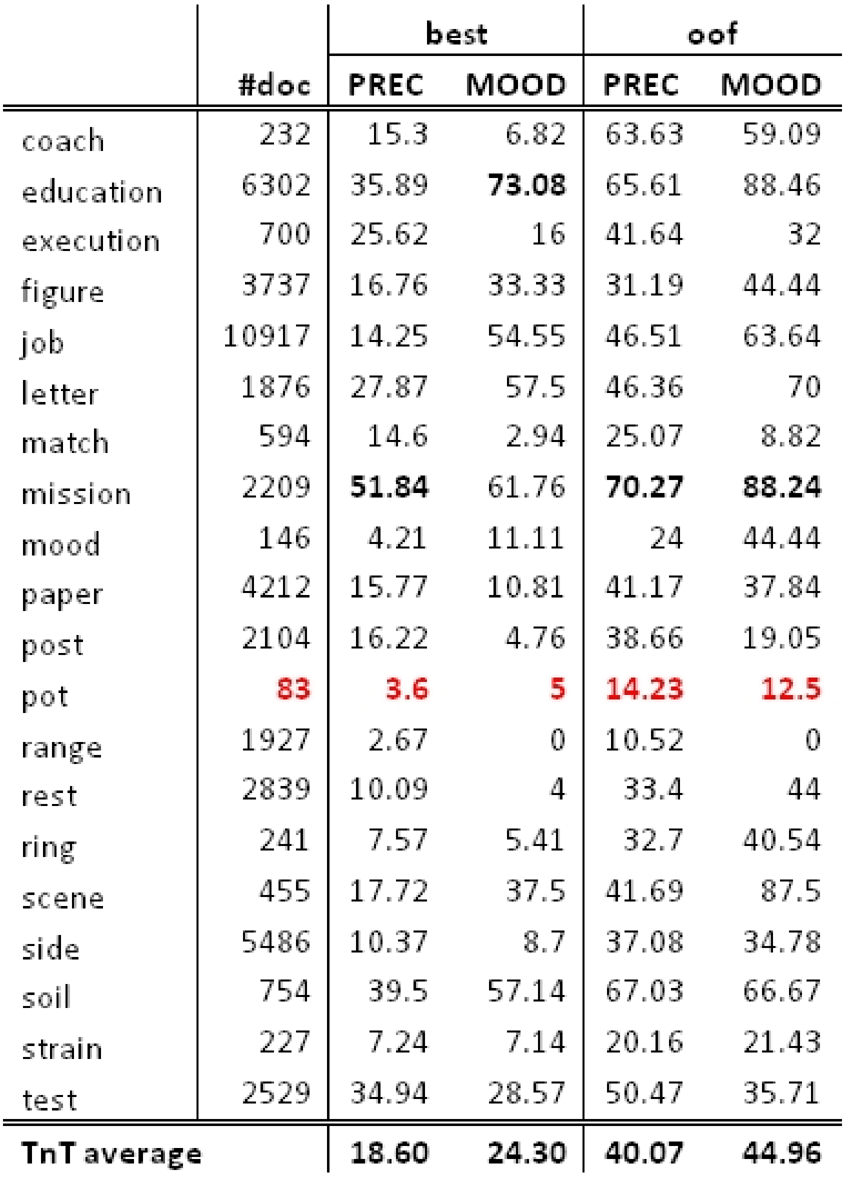
Appendix A presents the detailed CLWSD evaluations for the XLING system per query per language. The content of appendix A includes:

- CLWSD Evaluation for XLING_TnT system (German)
- CLWSD Evaluation for XLING_TnT system (Spanish)
- CLWSD Evaluation for XLING_TnT system (French)
- CLWSD Evaluation for XLING_TnT system (Italian)
- CLWSD Evaluation for XLING_TnT system (Dutch)
- CLWSD Evaluation for XLING_SnT system (Germanic: de,nl)
- CLWSD Evaluation for XLING_SnT system (Romance: es,fr,it)



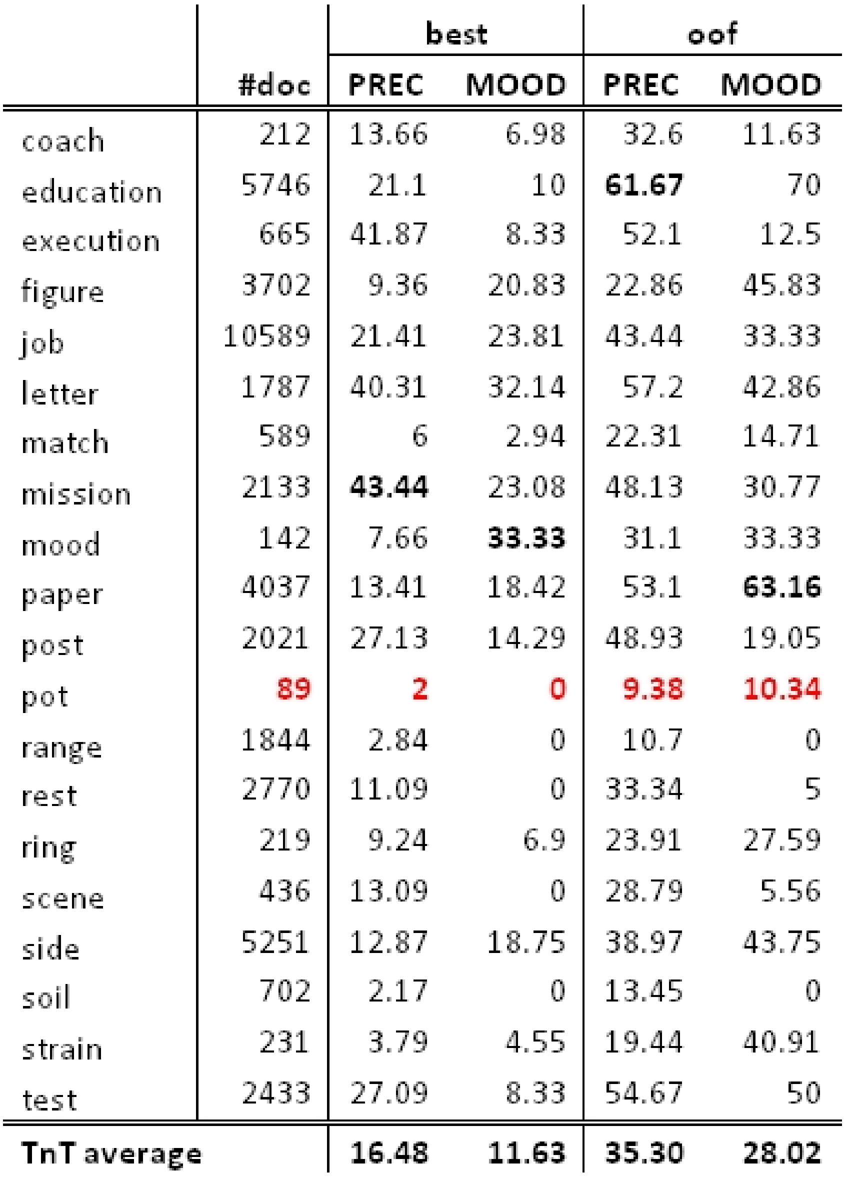
		best		oof	
	#doc	PREC	MOOD	PREC	MOOD
coach	294	1.67	0	5.42	0
education	6615	11.41	6.67	22.92	26.67
execution	735	5.87	5.88	25.5	29.41
figure	3830	1.63	0	5.99	0
job	11582	2.97	5.88	18.24	41.18
letter	1942	7.89	7.41	27.25	33.33
match	640	0.4	0	1.8	0
mission	2354	2.05	0	22.61	20
mood	160	**17.45**	0	**40.65**	8.33
paper	4468	5.9	5.88	25.87	29.41
post	2242	2.17	0	12.66	7.14
pot	**94**	**4.17**	**3.45**	**19.08**	**13.79**
range	2047	1.67	0	6.46	11.11
rest	2920	5.62	17.65	15.55	52.94
ring	249	3.78	0	19.27	15.63
scene	453	0.57	0	8.98	4.76
side	5515	11.25	13.64	41.02	54.55
soil	776	12.13	**33.33**	33.66	**95.83**
strain	238	1.86	0	9.47	10
test	2620	5.26	16.67	20.17	16.67
TnT average		**5.29**	**5.82**	**19.13**	**23.54**

Figure A.1: CLWSD Evaluation for XLING_TnT system (German)



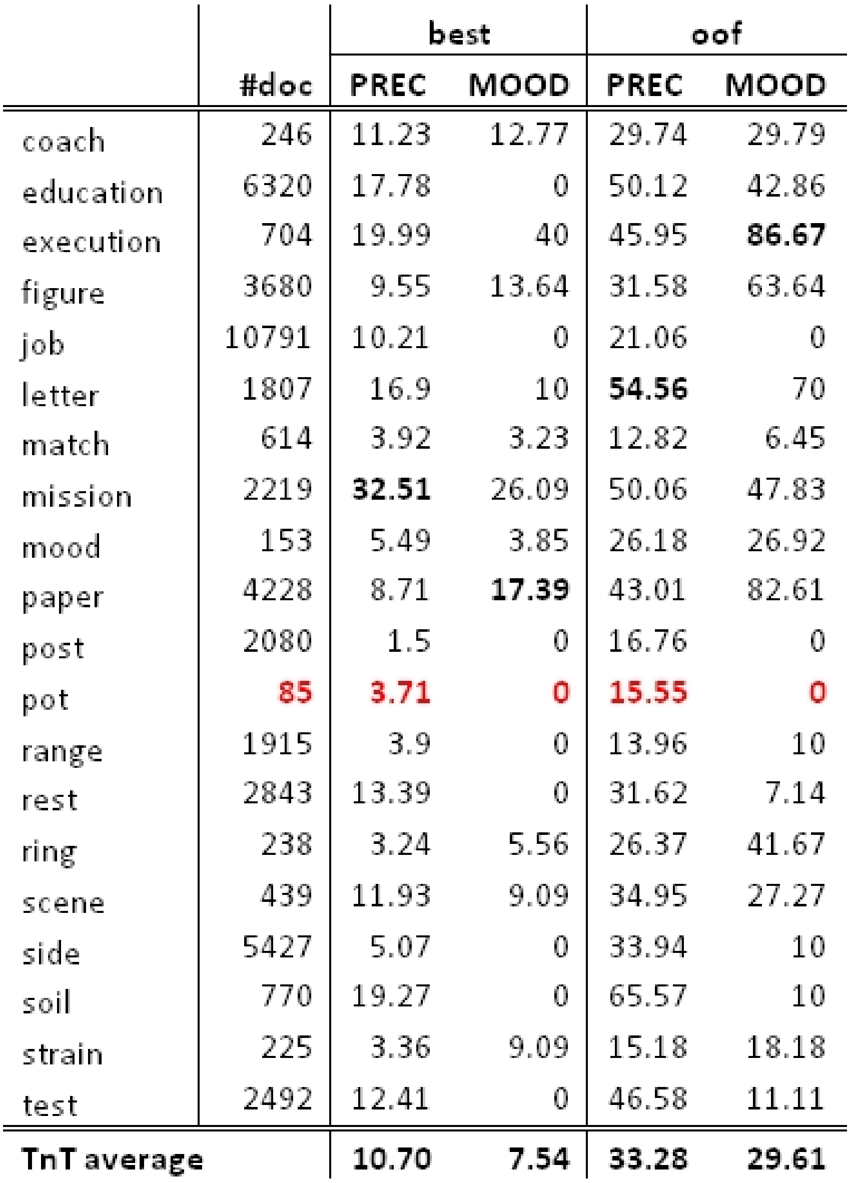
		best		oof	
	#doc	PREC	MOOD	PREC	MOOD
coach	232	15.3	6.82	63.63	59.09
education	6302	35.89	**73.08**	65.61	88.46
execution	700	25.62	16	41.64	32
figure	3737	16.76	33.33	31.19	44.44
job	10917	14.25	54.55	46.51	63.64
letter	1876	27.87	57.5	46.36	70
match	594	14.6	2.94	25.07	8.82
mission	2209	**51.84**	61.76	**70.27**	**88.24**
mood	146	4.21	11.11	24	44.44
paper	4212	15.77	10.81	41.17	37.84
post	2104	16.22	4.76	38.66	19.05
pot	**83**	**3.6**	**5**	**14.23**	**12.5**
range	1927	2.67	0	10.52	0
rest	2839	10.09	4	33.4	44
ring	241	7.57	5.41	32.7	40.54
scene	455	17.72	37.5	41.69	87.5
side	5486	10.37	8.7	37.08	34.78
soil	754	39.5	57.14	67.03	66.67
strain	227	7.24	7.14	20.16	21.43
test	2529	34.94	28.57	50.47	35.71
TnT average		**18.60**	**24.30**	**40.07**	**44.96**

Figure A.2: CLWSD Evaluation for XLING_TnT system (Spanish)



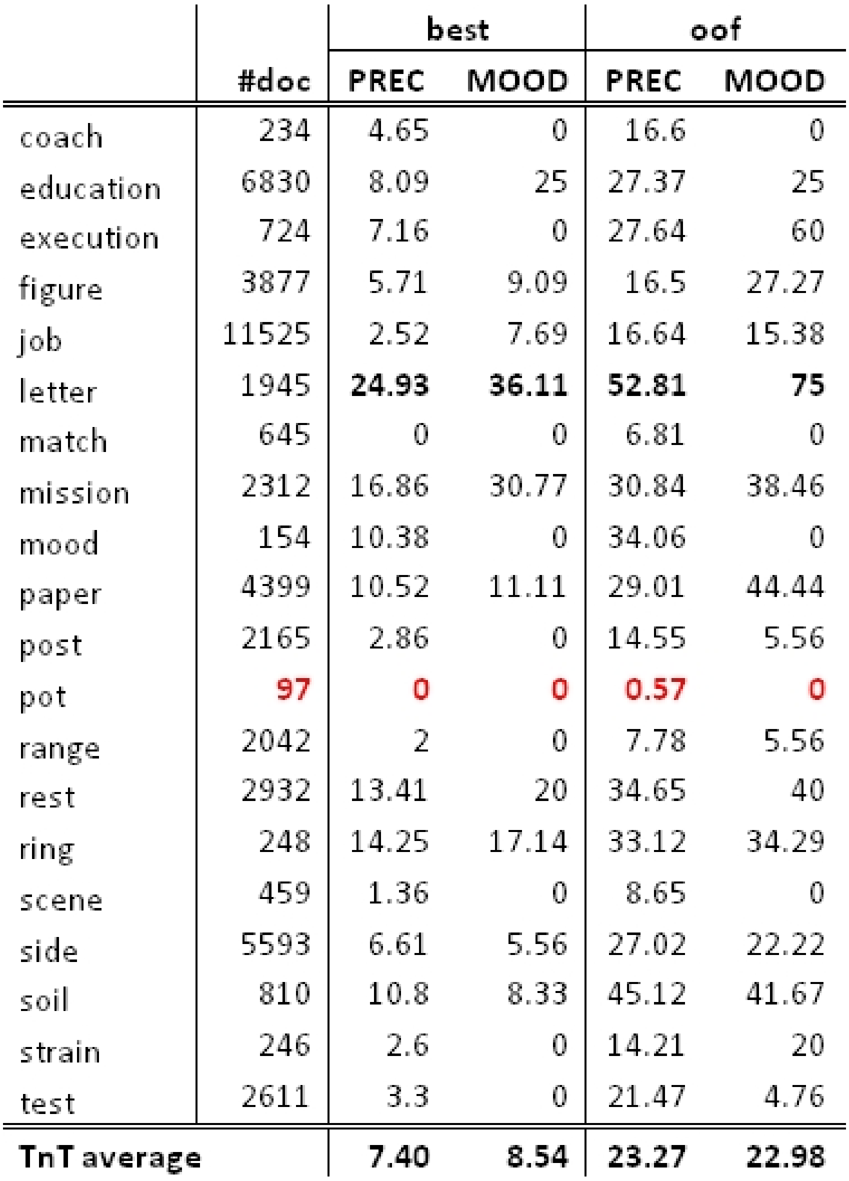
		best		oof	
	#doc	PREC	MOOD	PREC	MOOD
coach	212	13.66	6.98	32.6	11.63
education	5746	21.1	10	**61.67**	70
execution	665	41.87	8.33	52.1	12.5
figure	3702	9.36	20.83	22.86	45.83
job	10589	21.41	23.81	43.44	33.33
letter	1787	40.31	32.14	57.2	42.86
match	589	6	2.94	22.31	14.71
mission	2133	**43.44**	23.08	48.13	30.77
mood	142	7.66	**33.33**	31.1	33.33
paper	4037	13.41	18.42	53.1	**63.16**
post	2021	27.13	14.29	48.93	19.05
pot	**89**	**2**	**0**	**9.38**	**10.34**
range	1844	2.84	0	10.7	0
rest	2770	11.09	0	33.34	5
ring	219	9.24	6.9	23.91	27.59
scene	436	13.09	0	28.79	5.56
side	5251	12.87	18.75	38.97	43.75
soil	702	2.17	0	13.45	0
strain	231	3.79	4.55	19.44	40.91
test	2433	27.09	8.33	54.67	50
TnT average		**16.48**	**11.63**	**35.30**	**28.02**

Figure A.3: CLWSD Evaluation for `XLING_TnT` system (French)



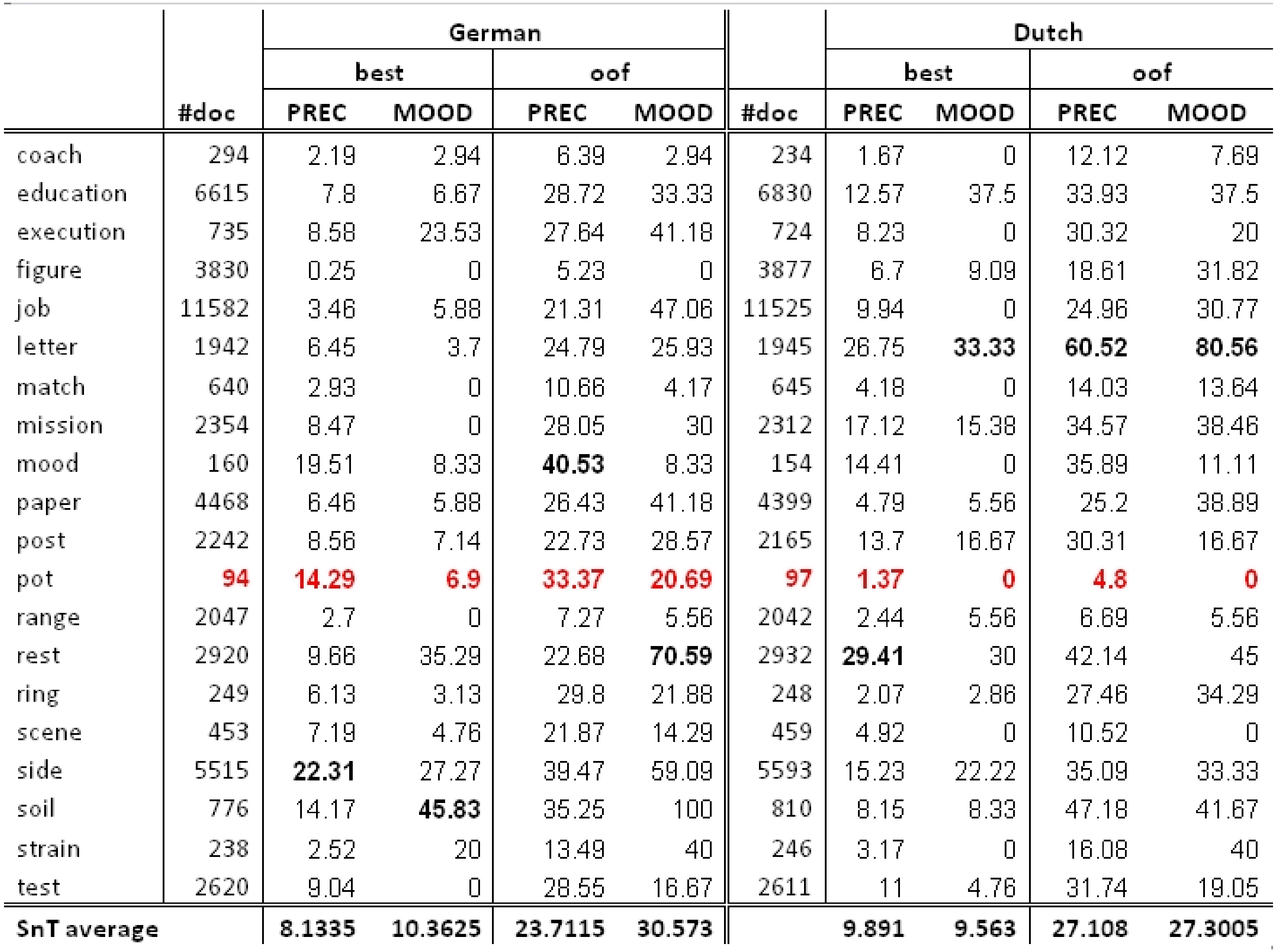
		best		oof	
	#doc	PREC	MOOD	PREC	MOOD
coach	246	11.23	12.77	29.74	29.79
education	6320	17.78	0	50.12	42.86
execution	704	19.99	40	45.95	**86.67**
figure	3680	9.55	13.64	31.58	63.64
job	10791	10.21	0	21.06	0
letter	1807	16.9	10	**54.56**	70
match	614	3.92	3.23	12.82	6.45
mission	2219	**32.51**	26.09	50.06	47.83
mood	153	5.49	3.85	26.18	26.92
paper	4228	8.71	**17.39**	43.01	82.61
post	2080	1.5	0	16.76	0
pot	**85**	**3.71**	**0**	**15.55**	**0**
range	1915	3.9	0	13.96	10
rest	2843	13.39	0	31.62	7.14
ring	238	3.24	5.56	26.37	41.67
scene	439	11.93	9.09	34.95	27.27
side	5427	5.07	0	33.94	10
soil	770	19.27	0	65.57	10
strain	225	3.36	9.09	15.18	18.18
test	2492	12.41	0	46.58	11.11
TnT average		**10.70**	**7.54**	**33.28**	**29.61**

Figure A.4: CLWSD Evaluation for `XLING_TnT` system (Italian)



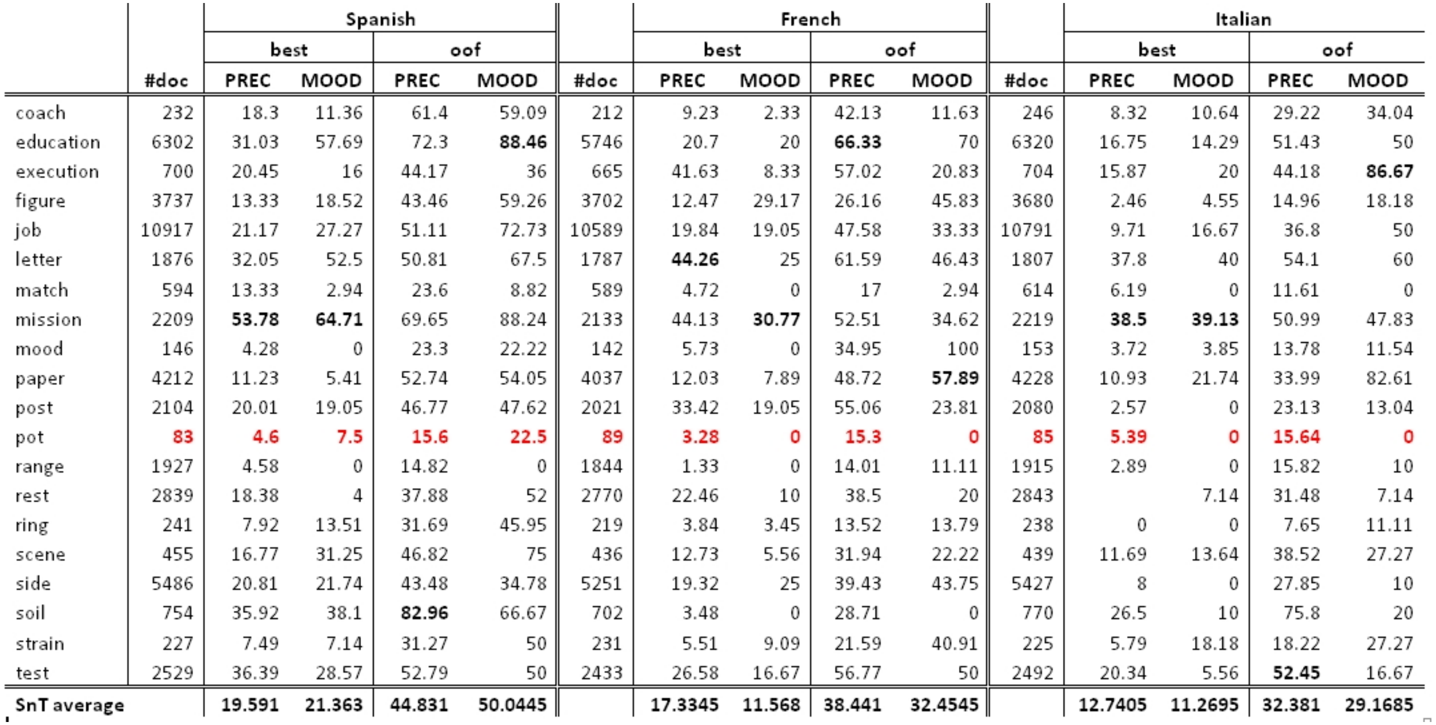
		best		oof	
	#doc	PREC	MOOD	PREC	MOOD
coach	234	4.65	0	16.6	0
education	6830	8.09	25	27.37	25
execution	724	7.16	0	27.64	60
figure	3877	5.71	9.09	16.5	27.27
job	11525	2.52	7.69	16.64	15.38
letter	1945	**24.93**	**36.11**	**52.81**	**75**
match	645	0	0	6.81	0
mission	2312	16.86	30.77	30.84	38.46
mood	154	10.38	0	34.06	0
paper	4399	10.52	11.11	29.01	44.44
post	2165	2.86	0	14.55	5.56
pot	**97**	**0**	**0**	**0.57**	**0**
range	2042	2	0	7.78	5.56
rest	2932	13.41	20	34.65	40
ring	248	14.25	17.14	33.12	34.29
scene	459	1.36	0	8.65	0
side	5593	6.61	5.56	27.02	22.22
soil	810	10.8	8.33	45.12	41.67
strain	246	2.6	0	14.21	20
test	2611	3.3	0	21.47	4.76
TnT average		**7.40**	**8.54**	**23.27**	**22.98**

Figure A.5: CLWSD Evaluation for `XLING_TnT` system (Dutch)



	German					Dutch				
		best		oof			best		oof	
	#doc	PREC	MOOD	PREC	MOOD	#doc	PREC	MOOD	PREC	MOOD
coach	294	2.19	2.94	6.39	2.94	234	1.67	0	12.12	7.69
education	6615	7.8	6.67	28.72	33.33	6830	12.57	37.5	33.93	37.5
execution	735	8.58	23.53	27.64	41.18	724	8.23	0	30.32	20
figure	3830	0.25	0	5.23	0	3877	6.7	9.09	18.61	31.82
job	11582	3.46	5.88	21.31	47.06	11525	9.94	0	24.96	30.77
letter	1942	6.45	3.7	24.79	25.93	1945	26.75	**33.33**	**60.52**	**80.56**
match	640	2.93	0	10.66	4.17	645	4.18	0	14.03	13.64
mission	2354	8.47	0	28.05	30	2312	17.12	15.38	34.57	38.46
mood	160	19.51	8.33	**40.53**	8.33	154	14.41	0	35.89	11.11
paper	4468	6.46	5.88	26.43	41.18	4399	4.79	5.56	25.2	38.89
post	2242	8.56	7.14	22.73	28.57	2165	13.7	16.67	30.31	16.67
pot	**94**	**14.29**	**6.9**	**33.37**	**20.69**	**97**	**1.37**	**0**	**4.8**	**0**
range	2047	2.7	0	7.27	5.56	2042	2.44	5.56	6.69	5.56
rest	2920	9.66	35.29	22.68	**70.59**	2932	**29.41**	30	42.14	45
ring	249	6.13	3.13	29.8	21.88	248	2.07	2.86	27.46	34.29
scene	453	7.19	4.76	21.87	14.29	459	4.92	0	10.52	0
side	5515	**22.31**	27.27	39.47	59.09	5593	15.23	22.22	35.09	33.33
soil	776	14.17	**45.83**	35.25	100	810	8.15	8.33	47.18	41.67
strain	238	2.52	20	13.49	40	246	3.17	0	16.08	40
test	2620	9.04	0	28.55	16.67	2611	11	4.76	31.74	19.05
SnT average		**8.1335**	**10.3625**	**23.7115**	**30.573**		**9.891**	**9.563**	**27.108**	**27.3005**

Figure A.6: CLWSD Evaluation for XLING_SnT system (Gothic)



		Spanish					French					Italian			
		best		oof			best		oof			best		oof	
	#doc	PREC	MOOD	PREC	MOOD	#doc	PREC	MOOD	PREC	MOOD	#doc	PREC	MOOD	PREC	MOOD
coach	232	18.3	11.36	61.4	59.09	212	9.23	2.33	42.13	11.63	246	8.32	10.64	29.22	34.04
education	6302	31.03	57.69	72.3	**88.46**	5746	20.7	20	**66.33**	70	6320	16.75	14.29	51.43	50
execution	700	20.45	16	44.17	36	665	41.63	8.33	57.02	20.83	704	15.87	20	44.18	**86.67**
figure	3737	13.33	18.52	43.46	59.26	3702	12.47	29.17	26.16	45.83	3680	2.46	4.55	14.96	18.18
job	10917	21.17	27.27	51.11	72.73	10589	19.84	19.05	47.58	33.33	10791	9.71	16.67	36.8	50
letter	1876	32.05	52.5	50.81	67.5	1787	**44.26**	25	61.59	46.43	1807	37.8	40	54.1	60
match	594	13.33	2.94	23.6	8.82	589	4.72	0	17	2.94	614	6.19	0	11.61	0
mission	2209	**53.78**	**64.71**	69.65	88.24	2133	44.13	**30.77**	52.51	34.62	2219	**38.5**	**39.13**	50.99	47.83
mood	146	4.28	0	23.3	22.22	142	5.73	0	34.95	100	153	3.72	3.85	13.78	11.54
paper	4212	11.23	5.41	52.74	54.05	4037	12.03	7.89	48.72	**57.89**	4228	10.93	21.74	33.99	82.61
post	2104	20.01	19.05	46.77	47.62	2021	33.42	19.05	55.06	23.81	2080	2.57	0	23.13	13.04
pot	**83**	**4.6**	**7.5**	**15.6**	**22.5**	**89**	**3.28**	**0**	**15.3**	**0**	**85**	**5.39**	**0**	**15.64**	**0**
range	1927	4.58	0	14.82	0	1844	1.33	0	14.01	11.11	1915	2.89	0	15.82	10
rest	2839	18.38	4	37.88	52	2770	22.46	10	38.5	20	2843		7.14	31.48	7.14
ring	241	7.92	13.51	31.69	45.95	219	3.84	3.45	13.52	13.79	238	0	0	7.65	11.11
scene	455	16.77	31.25	46.82	75	436	12.73	5.56	31.94	22.22	439	11.69	13.64	38.52	27.27
side	5486	20.81	21.74	43.48	34.78	5251	19.32	25	39.43	43.75	5427	8	0	27.85	10
soil	754	35.92	38.1	**82.96**	66.67	702	3.48	0	28.71	0	770	26.5	10	75.8	20
strain	227	7.49	7.14	31.27	50	231	5.51	9.09	21.59	40.91	225	5.79	18.18	18.22	27.27
test	2529	36.39	28.57	52.79	50	2433	26.58	16.67	56.77	50	2492	20.34	5.56	**52.45**	16.67
SnT average		**19.591**	**21.363**	**44.831**	**50.0445**		**17.3345**	**11.568**	**38.441**	**32.4545**		**12.7405**	**11.2695**	**32.381**	**29.1685**

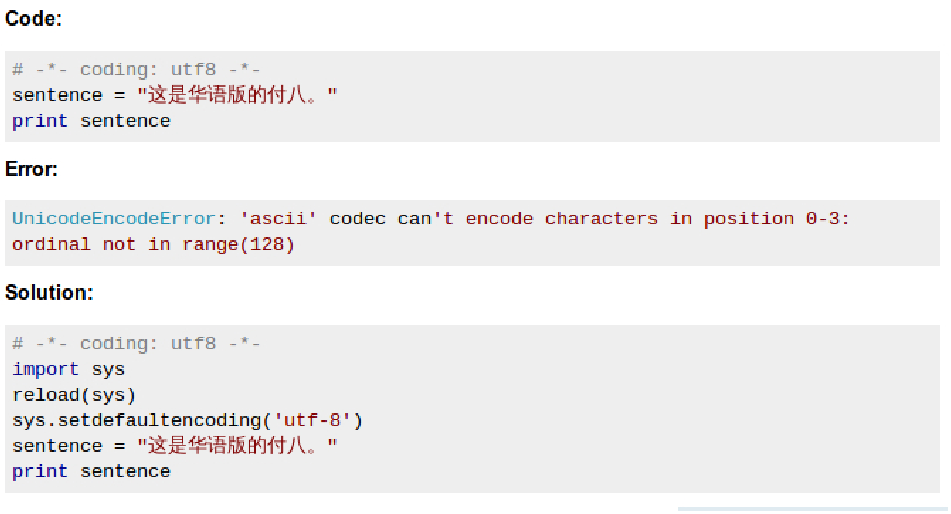
Figure A.7: CLWSD Evaluation for XLING_SnT system (Romance)

Appendix B

Appendix B documents the gotcha moments in coding in python for Multilingual Natural Language Processing (NLP). Often these problems are classified as too localized on online forums and hence they tend to be ignored. Solution documented in this appendix works for Python 2.7.4 and above but it does not work on Python 3.0+.

B.1 UTF-8 Processing in Python

As a convention to processing *Universal Character Set Transformation Format - 8 bits* (`utf-8`) textfiles, it is polite to inform the python interpreter to declare its python source code encoding to `utf-8`, otherwise the interpreter complains with errors when executing the code.[1] Declaration of soure code encoding can be done at the first line of the textfile. However, even if you have declared source code encoding at the start of your code, you might still encounter UnicodeEncodeError when you have unicode characters in your source code, e.g.:



Code:

```
# -*- coding: utf8 -*-
sentence = "这是华语版的付八。"
print sentence
```

Error:

```
UnicodeEncodeError: 'ascii' codec can't encode characters in position 0-3:
ordinal not in range(128)
```

Solution:

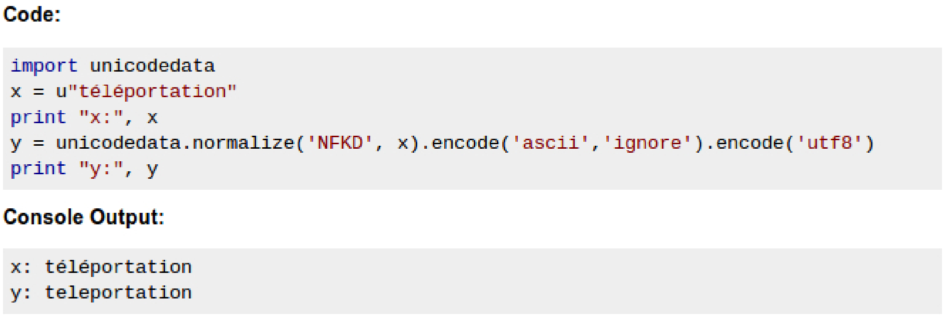
```
# -*- coding: utf8 -*-
import sys
reload(sys)
sys.setdefaultencoding('utf-8')
sentence = "这是华语版的付八。"
print sentence
```

This issue of setting default python encoding was discussed in depth at *Why we need sys.setdefaultencoding(utf-8) in a py script?*.

[1] for more detailed PEP, see `http://www.python.org/dev/peps/pep-0263/`

B.2 Normalizing Accented Latin Characters

Often in processing European languages we want to remove diacritics or accent marks from the tokens, for example, converting from *téléportation* to *teleportation*. More often than not, the original token with diacritics causes problem when it is fed as input to preprocessing tools. To normalize diacritic in python, you can use the `normalize` module in the `unicode` library as such:



Code:

```
import unicodedata
x = u"téléportation"
print "x:", x
y = unicodedata.normalize('NFKD', x).encode('ascii','ignore').encode('utf8')
print "y:", y
```

Console Output:

```
x: téléportation
y: teleportation
```

The option `NFKD` stands for *Normal Form Compatible Decompose*. By specifying the `NFKD`, python replaces all compatibility characters with their equivalents, e.g. accented `é` to `e`. For more details, see the official documentation from python API for `unicode.normalize()`.

Bibliography

Agirre, Eneko & German Rigau. 1996. Word sense disambiguation using conceptual density. In *Coling*, 16–22.

Agirre, Eneko & Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th conference of the european chapter of the acl (eacl 2009)*, 33–41. Athens, Greece: Association for Computational Linguistics. `http://www.aclweb.org/anthology/E09-1005`.

Anaya-Sánchez, Henry, Rafael Berlanga Llavori & Aurora Pons-Porrata. 2007. Retrieval of relevant concepts from a text collection. In *Caepia*, 21–30.

Anderson, J. R. & G. H. Bower. 1974. *Human associative memory.* Washington, DC: Hemisphere.

Anderson, John R. & Lael J. Schooler. 1991. Reflections of the environment in memory. *Psychological Science* 2(6). 396–408.

Balota, David A. & James I. Chumbley. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance* 10(3). 340 – 357.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bentivogli, Luisa, Pamela Forner & Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the MultiSemCor corpus. In *20th* *international conference on computational linguistics: Coling-2004*, 364–370. Geneva.

Bernard, John. 1987. *The macquarie thesaurus*. Macquarie Library. `http://books.google.com.sg/books?id=-noCHQAACAAJ`.

Blei, David M., Thomas L. Griffiths, Michael I. Jordan & Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *Nips*, .

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003b. Latent dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge & Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Sigmod '08: Proceedings of the 2008 acm sigmod international conference on management of data*, 1247–1250. New York, NY, USA: ACM. `http://ids.snu.ac.kr/w/images/9/98/SC17.pdf`.

Bond, Francis, Timothy Baldwin, Richard Fothergill & Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th global wordnet conference (gwc 2012)*, 56–63. Matsue.

Bond, Francis, Hitoshi Isahara, Kyoko Kanzaki & Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth international conference on language resources and evaluation (lrec 2008)*, Marrakech.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *In proceedings of the 6th global wordnet conference (gwc 2012). matsue.*, 6471.

Bond, Francis & Adam Pease. 2013. Linking and extending an open multilingual wordnet. In *In proceedings of 51st annual meeting of the association for computational linguistics*, .

Boyd-Graber, Jordan L., David M. Blei & Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Emnlp-conll*, 1024–1033.

Breen, James W. 2003. Word usage examples in an electronic dictionary. In *Papillon (multi-lingual dictionary) project workshop*, Sapporo. (`http://www.csse.monash.edu.au/~jwb/papillon/dicexamples.html`).

Buntine, Wray L. 2002. Variational extensions to em and multinomial pca. In *Ecml*, 23–34.

Buntine, Wray L. & Aleks Jakulin. 2004. Applying discrete pca in data analysis. In *Uai*, 59–66.

Burnard, Lou (ed.). 2007. *Reference Guide for the British National Corpus*. Research Technologies Service at Oxford University Computing Services.

Cai, Junfu, Wee Sun Lee & Yee Whye Teh. 2007. Improving word sense disambiguation using topic features. In *Emnlp-conll*, 1015–1023.

Charniak, Eugene. 2000. Bllip 1987-89 wsj corpus release 1. In *Linguistic data consortium*, Philadelphia.

Chklovski, Timothy & Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of the acl-02 workshop on word sense disambiguation: recent successes and future directions - volume 8* WSD '02, 116–122. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1118675.1118692. `http://dx.doi.org/10.3115/1118675.1118692`.

Collins, A. & M. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8(2). 240–247. doi:10.1016/s0022-5371(69)80069-1. `http://dx.doi.org/10.1016/s0022-5371(69)80069-1`.

de Cruys, Tim Van & Marianna Apidianaki. 2011. Latent semantic word sense induction and disambiguation. In *Acl*, 1476–1485.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer & Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41(6). 391–407.

Deese, James. 1959. On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology* 58(1). 17 – 22. `http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1960-03918-001&site=ehost-live`.

Diab, Mona & Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *40th annual meeting of the association for computational linguistics: Acl-2002*, 255262. Philadelphia.

Dolamic, Ljiljana & Jacques Savoy. 2010. When stopword lists make the difference. *J. Am. Soc. Inf. Sci. Technol.* 61(1). 200–203. doi:10.1002/asi.v61:1. `http://dx.doi.org/10.1002/asi.v61:1`.

Erk, Katrin & Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Emnlp*, 897–906.

Fellbaum, Christine (ed.). 1998. *WordNet: An electronic lexical database.* MIT Press.

Francis, W. N. & H. Kucera. 1964. *Brown corpus manual to accompany a standard corpus of present-day edited american english, for use with digital computers.* Brown University.

Francis, W. N. & H. Kucera. 1979. Brown corpus manual. Tech. rep. Department of Linguistics, Brown University, Providence, Rhode Island, US. `http://icame.uib.no/brown/bcm.html`.

Gale, William A. & Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1). 75–102.

van Gompel, Maarten. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th international workshop on semantic evaluation*, 238–241. Uppsala, Sweden: Association for Computational Linguistics. `http://www.aclweb.org/anthology/S10-1053`.

Griffiths, Thomas, Thomas L. Griffiths & Joshua B. Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological Science* 17. 767–773.

Griffiths, Thomas L. & Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101(Suppl 1). 5228–5235. doi:10.1073/pnas.0307752101. `http://www.pnas.org/content/101/suppl.1/5228.abstract`.

Griffiths, Thomas L., Joshua B. Tenenbaum & Mark Steyvers. 2007. Topics in semantic representation. *Psychological Review* 114. 2007.

Hawker, Tobias, Mary Gardiner & Andrew Bennetts. 2007. Practical queries of a massive n-gram database. In *Proceedings of the australasian language technology workshop 2007*, 40–48. Melbourne, Australia. http://www.aclweb.org/anthology/U07-1008.

Hoffman, Matthew D., David M. Blei & Francis Bach. 2010. Online learning for latent dirichlet allocation. In *In nips*, .

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Sigir*, 50–57.

Hoste, Véronique, Iris Hendrickx, Walter Daelemans & Antal van den Bosch. 2002. Parameter optimization for machine-learning of word sense disambiguation. *Natural Language Engineering* 8(4). 311–325.

Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama & Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth international conference on language resources and evaluation (lrec 2008)*, Marrakech.

Iyer, R.M. & M. Ostendorf. 1999. Modeling long distance dependence in language: topic mixtures versus dynamic cache models. *Speech and Audio Processing, IEEE Transactions on* 7(1). 30–39.

Ji, Heng & Ralph Grishman. 2011. Knowledge base population: successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies - volume 1* HLT '11, 1148–1158. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002472.2002618.

Jin, Peng, Yunfang Wu & Shiwen Yu. 2007. Semeval-2007 task 05: Multilingual chinese-english lexical sample. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, 19–23. Prague, Czech Republic: Association for Computational Linguistics. http://www.aclweb.org/anthology/S/S07/S07-1004.

Katz, Phil & Paul Goldsmith-Pinkham. 2006. Word sense disambiguation using latent semantic analysis. *Retreived from www.sccs.swarthmore.edu/users/07/pkatz1/cs65f06-final.pdf* .

Kilgarriff, Adam & Colin Yallop. 2000. What's in a thesaurus? In *Lrec*, .

Kintsch, Walter & W. Kintsch. 2001. Predication. *Cognitive Science* 25. 173–202.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Mt summit x*, .

Landauer, Thomas K & Susan T. Dutnais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 211–240.

Lesk, Michael. 1986. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 sigdoc conference*, 24–26. New York: ACM.

Li, Chenliang, Aixin Sun & Anwitaman Datta. 2011. A generalized method for word sense disambiguation based on wikipedia. In *Proceedings of the 33rd european conference on advances in information retrieval* ECIR'11, 653–664. Berlin, Heidelberg: Springer-Verlag. `http://dl.acm.org/citation.cfm?id=1996889.1996972`.

Li, Linlin, Benjamin Roth & Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Acl*, 1138–1147.

Lund, Kevin & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods* 28(2). 203–208. doi:10.3758/bf03204766. `http://dx.doi.org/10.3758/bf03204766`.

Lyons, J. 1977. *Semantics: 1* Semantics. Cambridge University Press.

Mahapatra, Lipta, Meera Mohan, Mitesh Khapra & Pushpak Bhattacharyya. 2010. Owns: Cross-lingual word sense disambiguation using weighted overlap counts and wordnet based similarity measures. In *Proceedings of the 5th international workshop on semantic evaluation*, 138–141. Uppsala, Sweden: Association for Computational Linguistics. `http://www.aclweb.org/anthology/S10-1028`.

Mallery, John C. 1988. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. In *Master's thesis, m.i.t. political science department*, .

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2008. *Introduction to information retrieval.* New York, NY, USA: Cambridge University Press.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY, USA: Henry Holt and Co., Inc.

McCarthy, Diana, Rob Koeling, Julie Weeds & John A. Carroll. 2004. Finding predominant word senses in untagged text. In *Acl*, 279–286.

McCarthy, Diana, Falmer East Sussex & Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *In proceedings of the 4th workshop on semantic evaluations (semeval-2007*, 48–53.

McDonald, Scott & Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd annual meeting on association for computational linguistics* ACL '04, Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1218955.1218958. `http://dx.doi.org/10.3115/1218955.1218958`.

Mihalcea, Rada. 2006. Random walks on text structures. In *Cicling*, 249–262.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3. 235–244.

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock & Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Hlt*, .

Mitchell, Jeff & Mirella Lapata. 2008. Vector-based models of semantic composition. In *Acl*, 236–244.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.* 41(2).

Navigli, Roberto & Simone Paolo Ponzetto. 2012a. Joining forces pays off: Multilingual joint word sense disambiguation. In *Emnlp-conll*, 1399–1410.

Navigli, Roberto & Simone Paolo Ponzetto. 2012b. Multilingual wsd with just a few lines of code: the babelnet api. In *Acl (system demonstrations)*, 67–72.

Navigli, Roberto & Paola Velardi. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(7). 1075–1086.

Ng, Hwee Tou & Hian Beng Lee. 1996a. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on association for computational linguistics* ACL '96, 40–47. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981863.981869. `http://dx.doi.org/10.3115/981863.981869`.

Ng, Hwee Tou & Hian Beng Lee. 1996b. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *In proceedings of the 34th annual meeting of the association for computational linguistics*, 40–47.

Niles, Ian & Adam Pease. 2001a. Towards a standard upper ontology. In Chris Welty & Barry Smith (eds.), *Proceedings of the 2nd international conference on formal ontology in information systems (fois-2001)*, Maine.

Niles, Ian & Adam Pease. 2001b. Towards a standard upper ontology. In *Proceedings of the international conference on formal ontology in information systems - volume 2001* FOIS '01, 2–9. New York, NY, USA: ACM. doi:10.1145/505168.505170. `http://doi.acm.org/10.1145/505168.505170`.

Och, Franz Josef & Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1). 19–51.

Pease, Adam & Christiane Fellbaum. 2010. *Ontologies and lexical resources* chap. Formal Ontology as Interlingua: The SUMO and WordNet Linking Project and GlobalWordNet. Cambridge: Cambridge University Press.

Proctor, P. 1978. *Longman dictionary of contemporary english.* Harlow, U.K.: Longman Group.

Rapp, Reinhard. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit*, `http://www.mt-archive.info/MTS-2003-Rapp.pdf`.

Roediger, H. L. & K. B. Mcdermott. 1995. Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21(4). 803–814.

Roediger, Henry L., Jason M. Watson, Kathleen B. McDermott & David A. Gallo. 2001. Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review* 8(3). 385–407.

Rogers, Timothy T. & James L. McClelland. 2008. Prcis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences* 31. 689–714. doi:10.1017/S0140525X0800589X. `http://journals.cambridge.org/article_S0140525X0800589X`.

Roget, P. M. 1852. *Roget's Thesaurus of English words and phrases.* Available from Project Gutemberg, Illinois Benedectine College, Lisle IL (USA). `http://www.amazon.co.uk/Rogets-Thesaurus-English-Words-Phrases/dp/0141004428`.

Salton, G. 1971. *The smart retrieval system&#8212;experiments in automatic document processing.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Silberer, Carina & Simone Paolo Ponzetto. 2010. Uhd: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th international workshop on semantic evaluation* SemEval '10, 134–137. Stroudsburg, PA, USA: Association for Computational Linguistics.

Singhal, Amit, Chris Buckley & Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th annual international acm sigir conference on research and development in information retrieval* SIGIR '96, 21–29. New York, NY, USA: ACM. doi:10.1145/243199.243206. `http://doi.acm.org/10.1145/243199.243206`.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* WWW '07, 697–706. New York, NY, USA:

ACM. doi:10.1145/1242572.1242667. `http://doi.acm.org/10.1145/1242572.1242667`.

Tan, Liling & Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proceedings of the 25th pacific asia conference on language, information and computation (paclic 25)*, 367–376. Singapore.

Tan, Liling & Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing* 22(3/4).

Thater, Stefan, Hagen Fürstenau & Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Ijcnlp*, 1134–1143.

Till, R. E., E. F. Mross & W. Kintsch. 1988. Time course of priming for associate and inference words in a discourse context. *Memory & Cognition* 16(4). 283–298+.

Tufiş, Dan, Dan Cristea & Sofia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology* 7(1–2). 9–34.

Turney, Peter D. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Ecml*, 491–502.

Turney, Peter D. 2006. Similarity of semantic relations. *Comput. Linguist.* 32(3). 379–416. doi:10.1162/coli.2006.32.3.379. `http://dx.doi.org/10.1162/coli.2006.32.3.379`.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)* 37. 141–188.

Ueda, Naonori & Kazumi Saito. 2006. Parametric mixture model for multitopic text. *Systems and Computers in Japan* 37(2). 56–66.

Ungerer, Friedrich & Hans-Jörg Schmid. 1996. *An introduction to cognitive linguistics*. Longman. `http://www.worldcat.org/isbn/9780582784963`.

Véronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language* 18(3). 223–252.

Vossen, Piek (ed.). 1998. *Eurowordnet: a multilingual database with lexical semantic networks.* Norwell, MA, USA: Kluwer Academic Publishers.

Wang, Yi, Hongjie Bai, Matt Stanton, Wen-Yen Chen & Edward Y. Chang. 2009. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Proceedings of the 5th international conference on algorithmic aspects in information and management* AAIM '09, 301–314. Berlin, Heidelberg: Springer-Verlag.

Xu, Renjie, Zhiqiang Gao, Yuzhong Qu & Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd asian semantic web conference (aswc 2008)*, 302–341.

Zipf, George Kingsley. 1949. *Human behaviour and the principle of least effort: an introduction to human ecology.* Addison-Wesley.