PROPOSAL FOR A PHD THESIS ON

SUPERVISED DOCUMENT TO ONTOLOGY INTERLINKING

by

Gabor Settimio Melli
M.Sc., SFU, 1998
B.Sc., UBC, 1989

THESIS PROPOSAL SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
School of Computing Science

SIMON FRASER UNIVERSITY

Spring 2010

ABSTRACT

We propose a thesis on the recognition of concept and relation mentions by supervised means in order to facilitate the interlinking of documents and ontologies. Concept mention identification will be performed with a trained CRF sequential model that enables the identification of mentions of concepts that are not yet be represented in the training corpus or ontology. Each mention will then associated with a set of candidate ontology concepts and a binary classifier used to predict the correct one. Finally, we propose the application of a supervised classifier to the identification of semantic relation mentions that can be added to the ontology.

The resulting system, SDOI, will be tested on a novel corpus and ontology from the field of data mining that we propose as a benchmark task. Some of the work has been completed and submitted for publication. Areas that remain to be explored include enhancements to the feature space of concept mention identification, and the integration of relation mention identification to the task.

Keywords: Concept Mention; Relation Mention; Reference Resolution; Ontology; Supervised Classification;

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1: INTRODUCTION

1.1 Motivation for the Thesis

The value from the growing availability in both electronic documents and ontologies will increase significantly once these two resources types have become deeply interlinked. Imagine for example the time when all the concepts and relations mentioned within a research paper are linked to the corresponding structure within a relevant ontology[1]; or when the concepts and relations mentioned in the business rules of a corporation's internal documents are linked to their corporate ontology. Documents will be grounded to formalized concepts. Ontologies will be grounded to the primary mechanism of human knowledge exchange.

The transition will facilitate navigational strategic reading of documents, when required. Further, searches on domain specific concepts such as "*supervised approaches to concept mention linking"* or "*exportable software product component*" could be satisfied more effectively than by the current ad-hoc approach of Web or corporate Intranet users having to iteratively fine-tune their keyword searches. Similarly, the use and development of ontologies will benefit from links to the usage in natural language. A concept or relation's meaning can also be more easily understood and improved upon by a person by seeing how it is used (described, constrained) in natural language. Finally, deep interlinking could enable new forms of information retrieval (Manning & al; 2008), information extraction (Sarawagi, 2008), topic modelling (Blei & Lafferty, 2007), document summarization (Melli & al, 2006), and Machine Reading (Etzioni & al, 2006).

An obstacle to this future of deeply interlinked information however is the significant amount of effort required from domain experts to insert the required additional information. Some automation of the linking step is a precondition to a future of deeply interlinked information, and some recent research suggests the large-scale feasibility of this automation by inductive means (Cucerzan, 2007; Mihalcea & Csamai, 2007; Milne & Witten, 2008; Kulkarni & al, 2009).

1.2 Proposed Topic and Approach

We propose a thesis on the design of a supervised algorithm for the task of concept and relation mention recognition with respect to an ontology. The task will be decomposed into three separate subtasks: 1) the identification of relevant concept mentions in a document; 2) the linking of each of these mentions to the appropriate concept in an ontology, if such a concept exists; and 3) the identification of relation mentions in a document that should be present in an ontology. This decomposition could support scenarios were one person first identifies the text segments that appear to refer to specialized meaning, then some other person with greater expertise could link the

[1] For example, the concept mentions in the abstract of this document are hyperlinked, or refer to http://www.gabormelli.com/RKB/2010_PhDThesisProposalSDOI.

mention to the correct concept in the ontology, and finally a third person with even greater expertise could identify relevant relations in the document.

Given the above decomposition, the proposed solution would first train a sequential classifier to identify token subsequences in a document as concept mentions. Motivations for the application of a sequential tagger include their successful use in the NLP community to the related tasks of text-chunking and named entity recognition (see Section3.2), and the possibility that any future improvements in the use and training of sequential taggers in other domain could be naturally imported into this framework. A further motivation of this approach is that it will identify token sequences that have not been encountered before, either elsewhere in the corpus or in the ontology.

Next, a binary supervised classifier will be applied to the concept mention linking task. Given the large number of classes (concepts in the ontology), each mention will be associated with a subset of candidate concepts by means of heuristic candidacy tests that can be understood as an informed method to undersample the data by removing cases that are very unlikely to be true. Next, each candidate concept is associated with a rich feature vector, including recursively defined (collective) features, and then labelled as true or false based on whether the concept is indeed the one that the mention must link to. In order to support the use of collective features we will investigate the effective use of an iterative supervised classifier that is simple enough to be reimplemented by other researchers.

Finally, another binary classifier would be applied to the task of relation mention identification. Each permutation of two (or possibly more) concept mentions would be associated with a feature vector and training label of whether the relation is present in the ontology.

The proposed solution will be implemented as a publicly available system (named `SDOI`) and be validated against a novel dataset consisting of the abstracts of the papers accepted to the KDD-2009 conference that has been linked to the concepts in a nascent data mining ontology.

Some of the content of the thesis will be based on work that has already been performed by the author, some of which has been published or submitted for publication (Melli & al, 2007; Melli & McQuinn 2008; Melli, 2010; Melli & al, 2010a; Melli & al; 2010b). The main original work that remains to be undertaken are in the areas of concept mention identification, and relation mention identification.

1.3 Intended Contributions

The main contributions intended for the thesis will be:

- A formal definition of the task of concept and relation mention recognition with respect to an ontology.
- A principled state-of-the-art supervised algorithm for the task that can realistically be reimplemented and extended.
- A novel and publicly available benchmark dataset that can be naturally expanded upon.

1.4 Proposed Timeline

The proposed time for delivery and defense of the thesis is as follows:

- March 16th – thesis proposal defense
- May 3rd – thesis submission
- June 1st - 4th – thesis defense

1.5 Outline of the Proposal

This thesis proposal is structured in a manner similar to that foreseen for the thesis. Sections 2 through 4 will define the task, present related work, and presents the evaluation dataset; next, sections 5 through 7 will describe the proposed algorithm. Specifically, section 5 will present the solution to the concept mention identification task; section 6 will present the solution to the concept mention linking task; and section 7 will present the solution to the relation identification task. Finally, section 8 will present the empirical evaluation, and section 9 will conclude the thesis and discuss future research directions.

2: TASK DEFINITION

We define the task of supervised concept and relation mention recognition with respect to an ontology in terms of its input requirements, output requirements, and performance measures.

2.1 Input Requirements

Assume that we are given a corpus of text documents $d_i \in D$ where each document is composed of sentences based on sequences of *tokens* (orthographic words or punctuation).

Assume also the existence of an ontology of interrelated *concepts*, $o_c \in O$, that represent and describe some concept within some domain. The concepts are interconnected by directed edges referred to as *internal links* (λ) that link one concept to another concept, $\lambda(o_{c'}, o_{c''})$. Each concept o_c can be associated with: a *preferred name*, p_c, a set of (also-known-as) synonyms A_c, and some descriptive text t_c. As described, an ontology is a directed and labelled multigraph that could be used to represent such diverse structures as Wikipedia[2] (with its rich text and weak semantics) to the Gene Ontology[3] (with its rich semantics and terse descriptions).

Assume next that each document d_i has a set of non-overlapping non-partitioning subsequences of tokens referred to as *concept mentions*, $m_m \in d_i$, that refers to a domain specific meaning not generally found in a dictionary. We assume that there is a significant overlap between the concepts intended for the ontology and the concepts mentioned in the corpus.

Every concept mention m_i is connected via a directed edge to either the concept o_j that captures the mention's intended meaning, or to the symbol "**?**" that denotes the absence of the concept within the ontology. We refer to these edges as *external links* and denote them as φ (m_m, o_c). An *unlinked* concept mention, φ (m_m, ?),is one that cannot be linked to the ontology because the concept is not yet deemed to be present in the ontology. We can refer to a mention's token sequence as its *anchor text*, a_m, to distinguish the text from the concept it links to.

Next, a *relation mention, r_i*, is a pairing of two concept mentions within the document $<m_{i',} m_{i''}>$. A relation mention is labelled as *true* if the mention signifies an internal link in the ontology, and *false* otherwise.

Figure 1 illustrates the concept mentions within a document. Next, Figure 2 illustrates the objects and relations available for analysis. Finally, Table 1 contains some additional terminology related to the task description.

[2] http://www.wikipedia.org

[3] http://www.geneontology.org

Figure 1 – An illustration of the task's training data. The two objects on top represent two text *documents*. The object below represents the *ontology* of *concept nodes* and *internal links*. Some non-overlapping subsequences in the documents are *concept mentions* mapped to either concept nodes or the unknown concept symbol (?), via *external links*. (This illustration will be updated to also illustrate all of the true relation mentions. Currently it only illustrates one relation mention with a dashed line.).

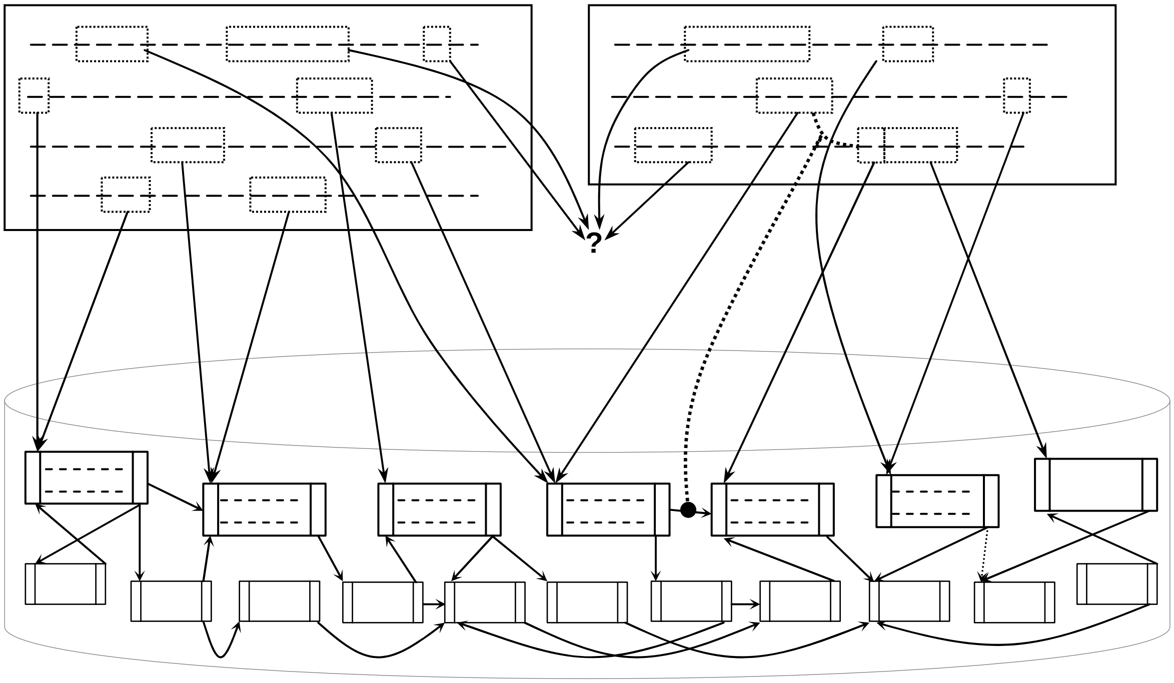

Figure 2 – The example of concept mention annotation using wiki-style formatting. Mentions are identified with doubled square brackets. The internal vertical bar (|) separates the *anchor text* from the concept reference. A question mark (?) refers to an unlinked concept.

[[*Collaborative Filtering Algorithm*| Collaborative filtering]] is the most popular [[*Algorithm*|approach]] to build [[Recommender System|recommender systems]] and has been successfully employed in many [[*Computer Application* |applications]]. However, as [[?|(Schein & al, 2002)]] explored, it cannot make recommendations for so-called [[?|cold start users]] that have rated only a very small number of [[*Recommendable Item*|items]].

Table 1 – Terminology associated with the task

m_i	The i^{th} concept mention in the corpus, $m_i \in D$.
o_i	The i^{th} concept node in the ontology, $o_i \in O$.
$I(o_i)$	The set of internal links into o_i from some o_k
$O(o_i)$	The set of internal links from o_i into some o_k.
$E(o_i, D)$	The set of external links into o_i from some $m_k \in D$.

2.2 Output Requirements

Given a document from the same domain as the ontology that lacks concept and relation mention information, the required output is the complete set of concept mentions within the document (both their *anchor text* and their corresponding *external link*), and the set of internal links mentioned in the document.

2.3 Evaluation

Several relevant evaluation criteria are available to measure performance. Concept mention identification can be naturally assessed in terms of F-measure (and its precision and recall components). An F-measure of 1.0 for example will occur when all concept mentions within a document are identified and no non-existing (false) mentions are predicted.

Concept mention linking is a multiclass classification problem that lends itself to measure of accuracy. An accuracy measure of 1.0 will occur when all true mentions within a document are linked to the correct concept node (one of which is the unlined concept symbol).

Relation mention identification is a binary classification task that can also be naturally assessed via an F-measure. On this subtask an F-measure of 1.0 will occur when all relation mentions between true concept mentions in a document that match a true internal link in the ontology are identified, and no non-existing (false) mentions are predicted.

Finally, the overall task can be assessed by measuring relation mention performance, not against true concept mentions, but against predicted concept mentions in the pipeline. As with the evaluation of the subtask then, the overall task can be measured in terms of F-measure. On the overall task an F-measure of 1.0 will occur when all relation mentions between predicted concept mentions in a document that are present in the ontology are identified, and no non-existing (false) mentions are predicted.

2.3.1 Partial Credit

A possible extension to the evaluation metrics that will be investigated is to grant partial credit for incorrect predictions that are relevant predictions. A motivating intuition for this extension is that the task can have a subjective aspect to it where even two people may not agree on exact assignments because ambiguities in language.

For the concept mention identification task, partial credit can be assigned if there is an overlap between the true mentions and the predicted mention. For the mention linking task partial credit can be assigned if the concept node selected only has a one edge distance to the correct node. For the relation mention identification task partial credit can be assigned if the relation is between concept mentions that only received partial credit.

2.3.2 Saving in Annotation Time

Given that the overall system performance will likely be below human-levels of performance, another measure that will be investigated is the proportion of time that would have been spent manually curating the information with assistance versus the amount of time spent working from pre-annotated content (which includes time for fixing mistakes made by the system). A proportion that is greater than 1.0 would indicate that some time was saved and would also suggest that the technology is ready for broad use. This human factors assessment assumes that the mechanisms implemented for annotating and for fixing annotations will not change significantly after more careful design of the data curation interfaces.

3: RELATED WORK

This chapter reviews some of the existing research that we expect to inform the proposed solution. The survey covers several topics from the fields of natural language processing, information retrieval, and information extraction, and also covers algorithmic topics drawn from data mining, machine learning and database research. Other related research areas that has been assessed but will not reviewed in detail include: 1) the extraction of technical terms to automatically create a book's subject index (Sclano & Velardi, 2007), 2) the automated population of ontologies (Builtelaar & al, 2008; Magnini & al; 2006), 3) database record deduplication (Bhattacharya & Getoor, 2004; Bilenko & al, 2005), and 4) unsupervised information extraction (Etzioni & al, 2008; Hassell & al, 2006).

3.1 Word Sense Disambiguation

One of the tasks within lexical semantics and natural language processing that resembles concept mention linking is that of *word sense disambiguation* (WSD) which requires that word mentions be linked to the appropriate word sense in a dictionary (Banerjee & Pedersen, 2002). Classic examples of the challenge are words such as "*bank*" and "*pike*", with their many wide-ranging senses. The Lesk algorithm is one of the more prevalent baseline algorithms for WSD (Lesk, 1986). The unsupervised algorithm uses a similarity measure that compares the overlap in the words in the dictionary definition and the words before/after the target word mention. The word sense with the highest overlap score is selected.

The task of WSD differs from our task in important ways. Solutions to the task can assume that the dictionary will contain most word senses in part because several electronic dictionaries exist with very broad lexical coverage. For concept mention linking on the other hand, one can assume that a large proportion of words will not be present in the ontology.

Further, in word sense disambiguation, the identification problem is trivial because dictionary words are typically composed of only one orthographic word and the lexical database is assumed to be complete. In our task the mentions can often be multi-word expressions and the word sense inventory (the ontology) is likely incomplete.

3.2 Named Entity Recognition

Another relevant task within lexical semantics is that of *named entity recognition* (NER) which requires the identification of proper names that refer to some set of basic entity types such as person, protein, organization, and/or location (Jijkoun, 2008; McCallum & Li, 2003). Examples of challenging names include words such as the location (island) of "Java", the company "Amazon", and the organization "the Ronald MacDonald charity".

While in named entity recognition the mentions can be complex multi-word expressions, the number of concepts to be linked to is significantly smaller than for the concept mention linking task which typically will not have a dominant concept type to link to.

3.3 Concept Mention Recognition in Biomedicine

A field that has actively investigated the ability to identify concepts in research papers and to link them to domain specific databases is the field of Biomedicine (Zweigenbaum & al, 2007). The focus of the field however, as in the BioCreAtIvE benchmark task (Crim & al, 2005), remains on identifying named entities, such as proteins, genes, and organisms; coping with the multitude of possible spellings and abbreviations; and then linking to the entity's specific database concept, such as the gene and protein database Swiss-Prot or Gene Ontology (Morgan & al, 2008).

3.4 Passage Linking

(Chakaravarthy *et al*, 2006) propose an unsupervised algorithm for linking sentence sequences in documents to database records. It makes use of TF-IDF-like ranking function that restricts itself to the terms that are available to describe entities. It proposes a greedy iterative cache refinement strategy to reduce the data retrieved from the entity database.

The algorithm requires that each segment be linked to at most one entity record. This restriction is acceptable for their scenario where each entity relates to one purchase transaction (because few segments will discuss more than one transaction) in our setting however can involve many mentions per single sentence.

3.5 Linking to Wikipedia and Wikipediafying

Recently some research has begun to investigate the more general task of identifying and linking of concept mentions, but restrict themselves to Wikipedia as the knowledge base (Cucerzan, 2007; Mihalcea & Csamai, 2007; Milne & Witten, 2008; Kulkarni & al, 2009). The approaches are generally tailored to take advantage of Wikipedia's internal structures such as category pages, disambiguation pages, and list pages, but can often be naturally extended to work against a more general defined ontology.

While Milne & Witten (2008) and Kulkarni & al (2009) focus on identifying and linking concept mentions within pages drawn from Wikipedia (Wikipediafying), their research also begins to explore the application on non-Wikipedia documents such as news articles. The two proposals also both apply some degree of supervised learning algorithms to the task, and both investigate features based on the relatedness of the document's overall topic to that of the concept's.

3.5.1 Milne & Witten, 2008

The most similar approach to the one that will be proposed for `SDOI` is that of (Milne & Witten, 2008). It proposes the use of a supervised classifier for linking mentions to Wikipedia that is based on three features. Two of the features are based on a proposed semantic relatedness measure between a candidate concept and the concepts mentioned in the document that are naturally disambiguable. As a first phase they detect the “context concepts” that do not require disambiguation, and then proceed to a supervised learning phase. For each mention they apply the rule of selecting the most confident link.

A challenge of applying their approach to the more general task of non-Wikipedia documents is their requirement that some of the mentions in the document can be naturally linked to the ontology without the need for disambiguation (in order to provide the context for the relatedness features). Their two phase approach could benefit from a more conservative approach that continually increases the mentions that will be deemed as disambiguated, and for these context concepts to be decided by the same trained classifier. Also, they use a very limited feature set in part because they conceived of the learning step mainly as a data-driven means to set a threshold on their three variables.

3.5.2 Kulkarni & al, 2009

(Kulkarni & al, 2009) extends the link selection work of (Milne & Witten, 2008) in two main ways. They add additional features based on the similarity between the bag-of-words representation of the text window surrounding the concept mention and the concept’s description in the ontology. They also propose a more sophisticated scheme to handle the relatedness-based collective features. Specifically they propose the use of an objective function that sums up the probability estimate produced by the trained classifier (based on bag-of-word features) and the relatedness measure proposed in (Milne & Witten, 2008) tested on all pairs of candidate concepts. They empirically show that optimizing on the proposed function closely tracks F1-measure performance (on several of their test documents, as the value of their objective function increases so did F1 performance). They explore two optimization algorithms for finding an optimal link assignment to the objective function. One algorithm is based on integer linear programming while the other based on greedy hill-climbing.

Foreseen challenges to the application of this approach to our task include its use of the longest matching sequence heuristic for concept mention identification which will reject many candidate mentions. Also, updating the proposed objective function to include additional features, or new definitions of relatedness, could unwittingly degrade algorithm performance.

3.6 Multiclass Classification

The general use of a supervised binary classification algorithm to solve multiclass tasks has been extensively researched and defended in the literature (Rifkin & Klautau, 2004; Fürnkranz, 2002). Two common approaches to the use of a binary classifier are to train a model for each class, OVA (Rifkin & Klautau, 2004), or to train a model for each pair of classes, AVA (Fürnkranz, 2002). Challenges to the application of these two generalized approaches to the task of concept and relation mention linking include:

1) Some classes in our test set will have few if any examples in the training set. The application of an OVA or AVA based approach would result in many concept mentions never being predicted.

2) The number of classes in our task is far larger than that tested in the literature. The dataset with the most classes tested by (Rifkin & Klautau, 2004) is the

spectrometer dataset from the UCI repository[4] with 48 classes. Our task could require thousands of classifiers to be trained which would likely result in a loss of performance because a large number of classifiers increase the chances that one of the classifiers will falsely claim to be the correct answer, and because we do not expect to have large amounts of training data per class.

3.7 Graph-Edge Prediction

Another relevant research area to the task is that of graph mining (Getoor & Diehl, 2005), specifically the prediction of whether two nodes are linked based on positive and negative examples of links. We focus the review on the work by (Al Hasan & al, 2006) that proposes the creation of feature vectors for each node pair and then apply a binary classification algorithm. Part of their contribution is an exploration of several different sources of information for predictive features ranging from topological features (such as the number of local edges) to intrinsic ones such as (matches between attribute-values). For the thesis we plan to reconcile their feature categorizations against the one we propose in section 6.3.

A notable difference in their task however is they do not require that nodes (in our case mentions) be linked to only one other node (in our case ontology nodes). In their scenario, such as in social networks, a node can have one or more links. Specifically, they test the algorithm on the ability to predict a future co-authorship relation. They also report that there is an inherent and significant skew of negative examples to positive examples and point to the literature on algorithms designed to handle classification in the presence of skewed data.

3.8 Classifiers for Skewed Data

Given the large number of ontology concepts, it is possible that research into handling skewed target attribute distributions will play a role. For example, if the problem is converted into a binary classification one then there could be an overwhelming number of negative examples relative to positive examples. Two approaches orthogonal approaches to handling these scenarios involve data processing remedies such as undersampling or oversampling (Chawla & al, 2004), and algorithmic remedies such as adjustments to kernel function definition (Wu & Chang, 2004). Of these two approaches, we propose the use of informed undersampling (of negative cases) in section 6.2.

[4] http://kdd.ics.uci.edu/

4: THE DMSWO1 AND KDD09CMA1 DATASETS

This section describes a novel real-world dataset created to evaluate the proposed solution. We describe the resource prior to the description of the proposed algorithm in order to draw examples from the resource in describing the algorithm. The dataset is composed of an ontology for the field of data mining, and an annotated corpus of research paper abstracts also from the field of data mining. To our knowledge, this is the first ontology and annotated corpus for a computing discipline. The dataset will be described in greater detail in the thesis, is the subject of a forthcoming paper (Melli, 2010a), and will be made publicly available[5].

4.1 The dmswo1 Data Mining Ontology

The dmswo1 ontology is based on a custom built semantic wiki[6] created specifically for the field of data mining and text mining by the author. In the wiki each concept has its own distinct page[7] and follows the structured English approach described in (Melli & McQuinn, 2008), where each concept contains: 1) A preferred name; 2) A one sentence definition in the form of "*an X is a type of Y that …*"; 3) A set of possible synonyms; 4) A set of relationships to other concepts stated in structured English; 5) A set of sample instances of the concept; 6) A set of counter-examples of the concept; 7) A set of related terms whose relationship has not been formally defined; and 8) a set of relevant external references for the concept. Table 2 summarizes some statistics of the ontology.

Table 2– Summary statistics of the dmswo1 ontology

CONCEPTS	4,659		
INTERNAL LINKS	25,170		
	MIN	MEDIAN	MAX
LINKS INTO A CONCEPT	0	3	157
LINKS OUT OF A CONCEPT	2	3	444
SYNONYMS PER CONCEPT	0	1	8

4.2 The kdd09cma1 Annotated Corpus

The author has also created an annotated corpus in order to evaluate, kdd09cma1, Additional motivations for the creation of the corpus include the lack of similar resources, and the possibility that this corpus could be the seed of a valuable and naturally expanding corpus.

[5] www.gabormelli.com/Projects/SDOI/v1.0/

[6] A semantic wiki is a wiki that captures semantic information in a controlled natural language that enables the generation of a formal machine-processable ontology http://www.semwiki.org/

[7] E.g. http://www.gabormelli.com/RKB/Information_Extraction_Task

The kdd09cma1 corpus is composed of the 139 abstracts for the papers accepted to ACM's SIGKDD conference, which took place in 2009 (KDD-2009)[8]. The competitive peer-reviewed conference on the topic of data mining and knowledge discovery from databases has acceptance rates in the range of 20% -25%. The annotation of the corpus (identification and linking of concept mentions) was performed in two separate phases. We first identified mentions of concepts that would be understood and/or often used within the data mining community. This phase was performed without consideration for what concepts existed in the ontology. Next, an attempt was made to link the mentions to the concept in the ontology (described in the next section) that stood for the intended concept in the mention. On average, the identification task took approximately 6 minutes per abstract, while the linking task took approximately 17 minutes per abstract. To evaluate the quality of the annotation, sixteen abstracts were randomly selected and the paper's author was asked to review the annotation. Fourteen authors responded and simply accepted the annotation as is.

The corpus bears similarities to corpora from the bio-medical domain such as the GENIA[9] and BioCreAtIvE[10] that are based on research paper abstracts found in MEDLINE abstracts and the terms are linked to concept in some ontology. Those corpora however focus on the annotation of basic named entities such as molecules, organisms, and locations. The kdd09cma1 corpus on the other hand contains very few named entities. Being from a formal science, its concept mentions range from single token ones such as "*mining*" to multi-token ones such as "*minimal biclique set cover problem*". Also, in cases where named entities do appear they often are embedded within an abstract concept mention, as in "*Gibbs sampling method*". The text was tokenized and assigned a part-of-speech role by using Charniak's parser [3]. Table 3 summarizes some key statistics about the corpus.

Table 3 – Summary statistics of the kdd09cma1 corpus, including the minimum, median, and maximum per abstract.

DOCUMENTS		139	PER DOCUMENT (min/med/max)
SENTENCES		1,185	3/8/17
TOKENS		30,450	105/220/367
CONCEPT MENTIONS	(100%)	5,449	25/57/95
SINGLE TOKEN	(~68%)	3,712	7/25/57
MULTI TOKEN	(~31%)	1,737	2/12/34

[8] The KDD-2009 abstracts are freely accessible from ACM's Digital Library http://portal.acm.org/toc.cfm?id=1557019

[9] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

[10] http://biocreative.sourceforge.net/

Given the novelty of the corpus and ontology, Table 4 summarizes some additional key statistics of the linking (external links) between the corpus and ontology.

Table 4 – Summary statistics of the external links from the `kdd09cma1` corpus to the `dmswo1` ontology.

DOCUMENTS	139	PER DOCUMENT (min/median/max)
LINKED MENTIONS	(69.0%) 3,758	10 / 25/ 59
UNLINKED MENTIONS	(31.0%) 1,691	2 / 12 / 31
DISTINCT CONCEPTS LINKED TO BY CORPUS	775	8 / 19 / 51
CONCEPTS UNIQUELY LINKED TO BY A SINGLE DOCUMENT		0 / 2 / 18

5: SUPERVISED CONCEPT MENTION IDENTIFICATION

To identify concept mentions in a document we plan to train a conditional random field (CRF) sequence tagging model in the same spirit as proposed in (Sha & Pereira, 2003) for the task of text chunking, and in (McCallum & Li, 2003) for and named entity recognition. The approach predicts first token of a mention (labelled with character B), any remaining mention tokens (labelled with I), and all other tokens are labelled with O.

These approaches generally make use of at two feature sources: the token itself, and its part of speech (POS) role. For the POS information, the use of an automated part-of-speech tagger (rather than manual annotation) is accepted practice. Figure 3 illustrates the labels used to identify concept mentions.

Figure 3 – Sample of the first sentence in Figure 1 labelled for concept mention identification.

```
Collaborative/B filtering/I is/O the/O most/O popular/O
approach/B to/O build/O recommender/B systems/I and/O has/O
been/O successfully/O employed/O in/O many/O applications/B
./O
```

As defined the tagger does not make direct use of the information in the ontology. A concept mention using a word that has not been previously encountered in the training data would likely not be identified. For the thesis we plan to extend the feature space by adding a feature that indicates the presence of the longest matching string in the ontology.

6: SUPERVISED CONCEPT MENTION LINKING

6.1 Approach

The approach proposed for supervised concept mention linking is to: 1) associate a set of candidate concepts for each mention, 2) associate a set of features with each candidate, 3) train a binary supervised classifier, and 4) apply a mention-level classifier to produce a prediction for each mention.

6.2 Candidate Concept Sets

A concept mention can be linked to any one of the many concept nodes in the ontology. However, knowledge of the mention's anchor text can be used to make an informed decision that will significantly reduce the number of concept nodes that should be realistically considered as candidates for the assignment, without discarding the correct node in the process. As an example, assume that a concept mention contains the anchor text composed of the single token of "*features*" then its candidate concept set might include the concepts for "Predictor Feature", "Application Feature", and "Data Table Attribute", one of which ideally is the correct concept.

This section explores a composite heuristic used to create the *candidate set* from a given anchor text, where a candidate set is composed of zero or more distinct concepts from the ontology: $a_m \rightarrow C_m = \{\varnothing, o_{c'}, o_{c''}, \ldots\}$. The heuristic is composed of a set of eight individual tests between a concept mention's anchor text and some information about a concept $t_i(a_m, o_c)$. Each test results in a set of accepted concepts. The overall heuristic accepts the union of all selected tests. Thus, a concept must pass at least one of the tests to become a member of a mention's candidate set. The actual set of tests proposed for SDOI will be determined empirically (see Section 8.4.1).

The first test to be considered, t_1, requires an exact match between the anchor text and the concept's preferred name. The second test, t_2, extends this pattern and requires that the anchor text exactly match any one of the concept's pre-identified synonyms (e.g. as materialized in the redirect pages in Wikipedia). These two tests can be used to replicate the proposals in (Milne & Witten, 2008; Kulkarni & al, 2009). In more specialized domains however, with complex multi-token mentions and with nascent ontologies that have small and incomplete synonym sets, these two tests would result in a weak recall rate of the correct concept. The anchor text of "*supervised learning of a sequential tagging model*" for example would be missed.

We define two additional candidacy tests for consideration. One of the tests, t_3, probes into the documents in the training corpus to determine whether the anchor text was also linked to this concept. In a sense, this test extends that of the synonym test in that at some future time some of these matching anchor texts will likely become official synonyms for the concept.

Given that a large proportion of concept mentions and concept synonyms are composed of more than one token, the final primary test, t_4, accepts a concept node where any of the component tokens match. Table 2 summarizes the four *primary* tests of candidacy.

Table 5 – the primary tests used to determine whether concept node (o_c) becomes a member of the candidate concept set (C_m) for anchor text (a_m).

t_1	The anchor text (a_{m}) matches the concept's preferred name (p_{c})
t_2	The anchor text (a_{m}) matches a synonym of the concept (o_{c})
t_3	The anchor text (a_{m}) matches a linked *anchor text* (in some other document) to the concept, $a_{k'} \in d_k$, $\varphi(a_{k'}, o_c)$ *and* $k \neq m$
t_4	A token in the anchor text (a_{m}) matches a token within the preferred name (p_{c}), a synonym (s in S_{c}), or a linked anchor text (a_{k}) in some other document

Finally, each of the four primary tests will be associated with an alternative test that is based on the use of the stemmed versions of the text being compared. We denote these tests as: t_{s1}, t_{s2}, t_{s3}, and t_{s4}. Note that, if a test succeeds on a primary test then it will also succeed on the stemmed version of the test.

6.3 Linking Features

Given a candidate concept set for each mention, and given training examples that identify the correct concept, the task of identifying the correct concept can be accomplished by training a supervised binary classification model. This section describes the feature vector associated with each paired mention/concept training case that will be used to train the classifier. Table 6 illustrates the structure of the training data produced.

Note that the features under the category of “Collective Features” are recursively defined on the classifiers labelling decisions. To handle the population of these features we perform iterative classification as described in Section 6.4.

Table 6 – Illustration of the structure of the training data used for the linking task.

Mention	Concept	Predictor Features													Label
		Anchor Txt.		Text Wind.	Document		Concept		Corpus		Cand. Set		Collective		
m_m	o_c	a_m,o_c		t_m,o_c	d_m,o_c		o_c		$D\text{-}d_m$		C_{mc}		o_c,S_m		T/F
1	1903	0	…	0.01	0.03	…	3	…	0	…	15	…	4	…	F
1	1021	0	…	0	0.01	…	1	…	2	…	30	…	1	…	F
1	829	1	…	0.02	0.02	…	12	…	1	…	15	…	7	…	T
2	4028	0	…	0.08	0.08	…	5	…	11	…	30	…	9	…	T
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

6.3.1 Anchor Text-based Features *f*(a_m,o_c)

Each of the six tests performed to determine concept candidacy are included as binary features. The intuition for their inclusion is that these tests can signal how closely a mention's anchor text matches text associated with the concept node.

6.3.2 Text Window-based Features *f*(t_m,o_c)

Another source of features can be based on text near the concept mention (its *text window*) and from the text used to describe the concept in the ontology. The feature included in `SDOI` that is based on this information is the cosine distance between the normalized bag-of-word vector representations of the text window and of the ontology description. This feature is proposed in [4] (they also include dot product and Jaccard similarity).

6.3.3 Document-based Features *f*(d_m,o_c)

The entire document can be used to inform the classification decision. Two proposed features are: 1) the cosine distance between the normalized bag-of-word vector representations of the document and the ontology description (also proposed in [4]), and 2) the token position of the concept mention within the document (1st token, 2nd token …). The intuition of the later feature is that different types of concepts are expressed near the beginning of a document rather than later on.

FEATURE	DEFINITION
cos(d_m,o_c)	The bag-of-word cosine similarity between the document and the concept description.
tok(a_m,d_m)	Number of tokens between the start of the document and the first token in the mention.

6.3.4 Concept-based Features $f(o_c)$

The candidate concept node on its own along with its role within the ontology (without knowledge of the specific concept mention being considered) can also inform the classification decision. For example, [4] proposes the use of the frequency that a concept is linked to (its inlink count) as a feature. We include this count as a feature, $CI(o_c)$ and also include the count of internal links extending out of the concept, $CO(o_c)$. The first feature signals the popularity of the concept as a reference. The second feature can signal whether the concept has received significant attention by the ontology designers in the form of additional links. Table 7 summarizes these three features.

Table 7 – Candidate concept-based features

FEATURE	**DEFINITION** (see Table 1)
$CI(o_c)$	Cardinality of all internal links into o_c, i.e. $\|(o_c)\|$
$CO(o_c)$	Cardinality of all internal links from o_c, i.e. $\|O(o_c)\|$

6.3.5 Corpus Based-based Features $f(o_c,D)$

Interestingly, the training corpus can also be an alternative source of information for predictor features. Indeed the t_3 candidacy test is signals the presence in the corpus of an identical mention. An additional corpus-based feature, CE, is the count of documents that also have external links to the ontology concept (Table 8).

An implementational challenge with this source of features is that when calculating a mention's corpus-based feature care must be taken to occlude all of the other mentions that occur within the document. If an anchor text for example is repeated elsewhere in the same document then in order to truthfully replicate the testing environment, the other mentions in the document cannot influence the calculation of the feature. Thus, the value associated with the CE (a_m, o_c, D) feature for a given mention can differ for every document in the training corpus.

Table 8 – Corpus-based features

FEATURE	**DEFINITION** (see Table 1)
CE (a_m, o_c, D)	Cardinality of all external links into o_c, i.e. $\|E(o_c, D')\|$, where $d_m \notin D'$

6.3.6 Candidate Set-based Features $f(C_{mc})$

Awareness of the size and membership of entire set of candidate concepts associated with the concept mention can inform the classification decision. For example, it is riskier to pick a concept from a large candidate set than from a candidate set composed of only two members. Table 9 summarizes these features.

Table 9 – Candidate Concept Set-based Features

FEATURE	DEFINITION
$CC(C_i)$	Cardinality of the set of candidate concepts. i.e. $\|C_i\|$
$\Sigma CI(C_i)$	Count of internal links into all candidate concept nodes. $CI(o_{j'}) + CI(o_{j''}) + \ldots$, for all $o_j \in C_i$
$RCI(o_j,C_i)$	Relative proportion of the internal links into the candidate concept relative to overall size. $CI(o_j) / \Sigma CI(C_i)$
$\Sigma CO(C_i)$	Count of internal links out from all candidate concept nodes. $CI(o_{j'}) + CI(o_{j''}) + \ldots$, for all $o_j \in C_i$
$RCO(o_j,C_i)$	Relative proportion of the internal links out from the candidate concept relative to overall size. $CI(o_j) / \Sigma CO(C_i)$
$\Sigma CE(C_i)$	Count of external links into all candidate concept nodes. $CE(o_{j'}) + CE(o_{j''}) + \ldots$, for all $o_j \in C_i$
$RCE(o_j,C_i)$	Relative proportion of the external links into the candidate concept relative to overall size. $CE(o_j) / \Sigma CE(C_i)$

6.3.7 Collective-based Features $f(o_c,S_m)$

We describe a set of features whose calculation requires knowledge about the label for some of the links that we are trying to predict. With possession of some disambiguated links to the ontology, the ontology can be used to provide some background knowledge into the classification decision for the remaining links. For example, if we knew that a document mentioned the concept "*supervised learning algorithm*" then the decision of which candidate concept to predict for the mention of "*feature*" may be improved (i.e. *predictor feature*, not *computer program feature*, nor *database attribute*). How to attain such a partial set of labelled links will be addressed in the next section. Let S_m be the context set of disambiguated concepts in document d_m.

In order to replicate the work in (Milne & Witten, 2008) we include the relatedness measure they propose, which in turn is based on the Normalized Google Distance (*NGD*) metric (Cilibrasi & Vitanyi, 2007) that assesses the dissimilarity between two sets.

$$NGD(A,B) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|O|) - \log(\min(|A|,|B|))}$$

Given two ontology concept nodes (o_a, o_b) the relatedness function proposed in [9] tests the links into two concepts nodes(o_a, o_b), where A=$I(o_a)$, and B=$I(o_a)$. Also, the [0,∞] function range of *NDG* is converted to a similarity metric by truncating the output to [0,1] and then subtracting it by 1.

$$MW08rel(o_a,o_b) = 1 - NGD(o_a,o_b), \\ if\ NGD(o_a,o_b) < 1 \qquad else\ = 1$$

We extend this feature space by also including the components used in the calculation of *NDG*. We also include a Jaccard set similarity feature. Table 10 defines some relevant functions to collective analysis, while Table 11 defines the collective features. In the case where the context set of mentions is empty, such as initially when no mentions have been linked, these features all calculate to zero (0).

Table 10 – Functions used to define collective features based on the relatedness of two concept nodes.

FUNCTION	DEFINITION
Ca$\cap$b(o_j, o_k)	The cardinality of I(o_j) $\cap$ I(o_k)
Ca$\cup$b(o_j, o_k)	The cardinality of I(o_j) $\cup$ I(o_k)
max_ab(o_j, o_k)	The larger of |I(o_j)| or |I(o_k)|.
min_ab(o_j, o_k)	The smaller of |I(o_j)| or |I(o_k)|.
AMW08rel(o_j, O')	The average relatedness. of concept o_{j} and O', where $o_{\mathrm{j}} \notin O'$.

Table 11 – Definition of the collective features.

FEATURE	DEFINITION
$CS(S_i)$	The cardinality of the set of context concepts. i.e. $\|S_i\|$.
$\Sigma IS(S_i)$	The count of internal links to the set of context concepts.
$AIS(S_i)$	The average number internal links into each of the context concepts.
$AM\cap S(o_j,S_i)$	The average cardinality of the intersection between the links into concept o_j and the internal links into the anchor concepts in S_i. *Ca*$\cap$b(o_j,S_i), where $o_k \in S_i$ and $o_j \neq o_k$.
AMW08rel(o_j,S_i)	The average weighted relatedness between the concept node o_j and each of the concept nodes in S_i, as proposed in [9].
Σ*MW08* (S_i)	The sum of the relatedness between each concept in S_i. to the other concepts in S_i. This feature is proposed in [9] to inform the classifier about the entire context set.
AJacc(o_j,S_i)	The average Jaccard set similarity between the links into the o_j concept and each of the concept nodes in S_i.

6.4 Collective Feature Handling via Iterative Classification

The “Collective-based” features defined in Section 5.7 require that some portion of a document’s concept mentions be already linked to the ontology. (Milken & Witten,

2008) accomplish this assignment by first identifying some mentions heuristically as “context” mentions that do not require disambiguation. (Kulkarni & al, 2009) accomplish this assignment by specifying a custom objective function that is then optimized by, for example, greedily committing to the next highest scoring mention.

We propose an incremental approach, but one that is directed by a supervised learning algorithm. We accomplish this by applying an iterative classifier algorithm inspired by the one proposed in (Milken & Witten, 2008) that first trains a model on an idealized context set, S_m, all the correctly labelled mentions, and then during the testing phase iteratively grows the context set based on an increasing proportion of the most likely predictions.

Given that our collective features all have the value zero (0) initially, we enhance the approach by first training a model on all but the collective features to seed the first guesses with informed choices. Assume that we define a constant number of iterations μ and a set of N training instances. The proposed algorithm is presented in Figure 4.

Figure 4 – Proposed iterative classification algorithm.

```
1. Train model ($M_{\cancel{\text{col}}}$) without the collective features
2. Train a model ($M_{\text{col}}$) with the collective features
3. For each iteration of $\iota$ from 1 to $\mu$
   a. Calculate the value for the collective features
   b. Apply model $M_{\cancel{\text{col}}}$ to the test set, if $\iota$ is 1
      otherwise, apply model $M_{\text{col}}$ to the test set.
   c. Select the $\kappa$ most probable links, where $\kappa = N(\iota/\mu)$.
4. Output the final set of predictions on all mentions.
```

6.5 Mention-level Classification

Recall that the goal of the task is to select at most one concept per candidate set (one external link per mention). The classifier however may assign the label of “True” to more than one concept associated to a mention. When this is the case, `SDOI` uses a tie-breaking rule. A possible tie-breaking rule is to make a random selection. However, if the supervised classifier used also reports a value that can rank the predictions according to their likelihood (e.g. support vector machines, decision tree, and logistic regression) then `SDOI` uses this number to select the more likely concept.

Given a binary classifier that predicts true or false on whether a concept should be linked to a mention, what concept should be selected when two or more concepts are predicted to be true? The challenge arises because of the use conversion a multiclass task was encoded as a binary one and the resulting information needs to be decoded.

One solution for resolving these outcomes is to introduce a heuristic classification rule to pick between the options. One such rule can be to randomly pick one of the concepts. A more informed decision is to pick the more commonly linked-to concept. Finally, if the binary classifier selected associated a confidence value with each of its predictions then the heuristic can be to: pick the prediction with the highest confidence as proposed by (Milne & Witten, 2008) and by the multi-class classification literature (Rifkin & Klautau, 2004).

None of these rules however account for dependencies that may between the candidate concepts. In the scenario with the five candidate concepts, two of which are predicted to be true, imagine further than neither is associated with a significantly larger confidence and the one with the lower confidence has a much more commonly linked to. In these situations it may be more reasonable to pick the candidate with the lower confidence. Further, it may occasionally be better to pick the most common class (which generally is the unlinked label).

But how to determine a more nuanced and accurate rule? A possible mechanism is to use a data-driven approach that trains a classifier at the mention level based on aggregated features from the candidate set with the classification label being whether the original rule is correct or not. This approach has the further appeal that it approximates the ideal of modeling at the mention level – except that it accomplishes it in two phases.

Predictor features for the modeling at the mention level could include:

- TruePreds: The number candidates that are predicted as true
- FalsePreds: The number candidates that are predicted as false
- HighTrueConf: The highest confidence value for a true prediction (if one existed).
- HighFalseConf: The highest confidence value for a false predictions (if one existed).

The label associated with each mention is whether the default rule of picking the concept with the highest confidence is correct or not. Preliminary experimentation with decision trees produced the two trees below:

Model 1

```
HighTrueConf > 0.01 : CORRECT
HighTrueConf <= 0.01 : INCORRECT => UNLINKED
```

Model 2

```
HighTrueConf > 0.93618 : CORRECT
HighTrueConf <= 0.93618 :
|   HighTrueConf > 0.733662 :
|   |   FalsePreds <= 0 : INCORRECT
|   |   FalsePreds > 0 : CORRECT
|   HighTrueConf <= 0.733662 :
|   |   HighFalseConf <= -0.557562 :
|   |   |   FalsePreds <= 4 : INCORRECT
|   |   |   FalsePreds > 4 : CORRECT
|   |   HighFalseConf > -0.557562 :  INCORRECT
```

The production of a decision tree is encouraging because it suggests that a more nuanced classification rule may be possible. A review of the features used by the models suggest the more predictive feature is the confidence of the highest true prediction. However, once this value dropped to a certain level then other features become predictive - particularly on the false side: the number of false predictions and the value of highest false prediction. Counter-intuitively, sometimes having few associated false predictions false results in a less accurate prediction.

7: RELATION MENTION DETECTION

Finally, the proposed solution to the task of detecting relation mentions present in the ontology is to apply the supervised relation mention detector, TeGRR, proposed in (Melli & al, 2007). The motivation for this application is that the algorithm is extensible to the detection of relations mentions with possibly more than two constituent concept mentions, and where the concept mentions cross sentence boundaries.

7.1 Text Graph Representation

The TeGRR algorithm is based on a graph representation of a document. The text graph representation is composed of the following types of nodes and edges: 1) Intrasentential nodes and edges; 2) Sentence to Sentence edges; and 3) Coreference nodes and edges. A sample text graph, which makes use of the three edge types, is presented in Figure 5.

7.1.1 Intrasentential Nodes and Edges

Intrasentential nodes and edges are intended to represent the information contained within each sentence. Many candidates for these edges exist in the literature. They include word-to-word edges (Freitag & McCallum, 1999), shallow parsing edges (Zelenko *& al*, 2003), dependency parse tree edges (Sushanek *& al*, 2006), and phrase-structure parse tree edges (Zhang*& al*, 2006).

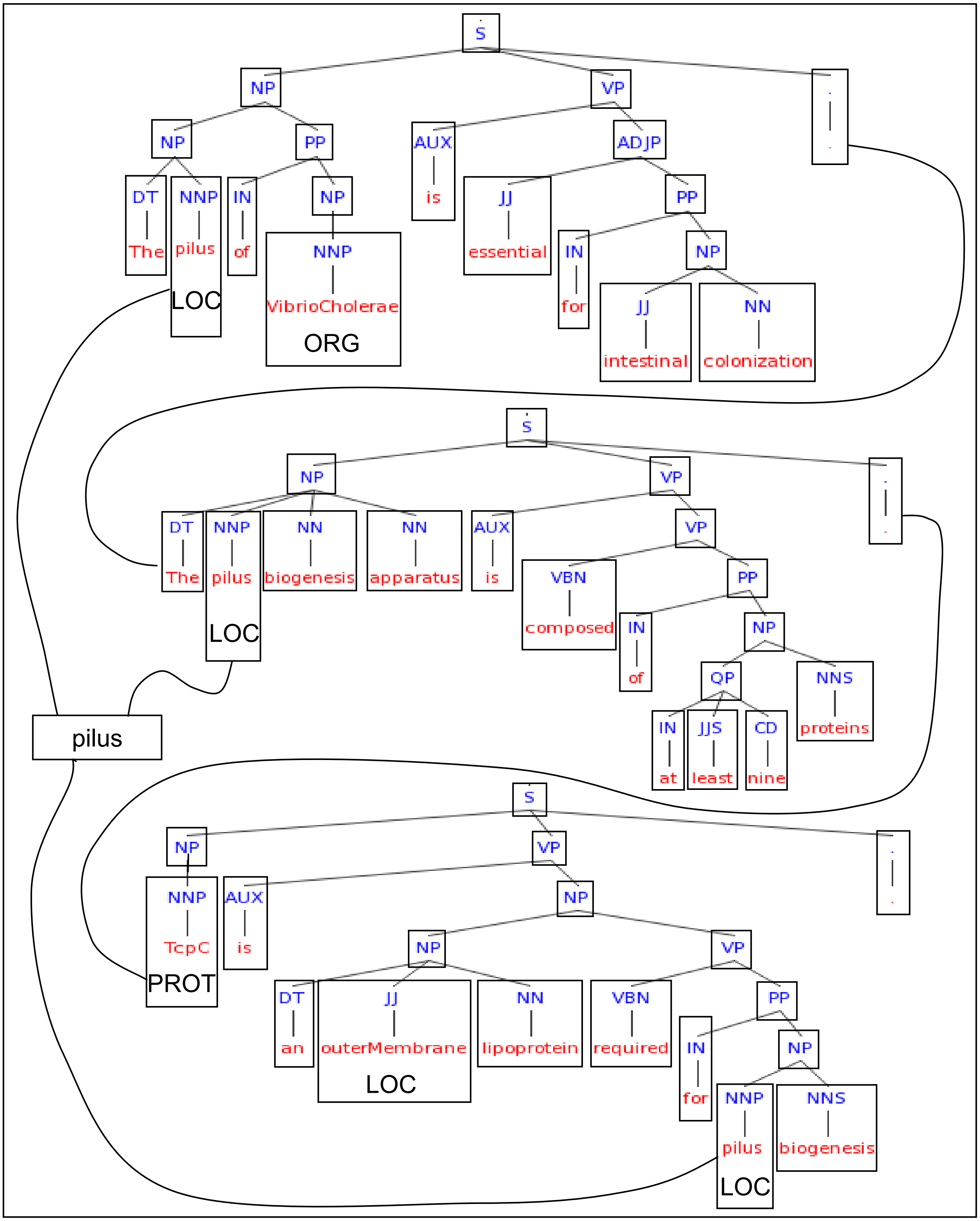

Figure 5 - A sample text graph representation used for relation mention identification (in the thesis this example will be drawn from the kdd09cma1 corpus)

We propose the use the phrase-structure parse tree as the source of intrasentential edges for two reasons. First, the recent analysis by (Jiang & Zhai, 2007) suggests that the phrase-structure parse tree is the single best source of information for relation detection. Secondly, all other proposed intrasentential edges can be derived from phrase-structure parse trees by means of simple transformations. Two types of nodes are associated to a phrase-structure parse tree: leaf nodes and internal nodes. Leaf nodes contain 1) a word, punctuation mark or entity instance, 2) the part of speech tag, and 3) a named entity tag if one exists. Internal nodes contain the syntactic phrase-structure label. The text graph in Figure 1 contains 52 intrasentential edges connecting 24 internal nodes and 32 leaf nodes.

7.1.2 Sentence-to-sentence Edges

The first type of intersentential edges considered is the “sentence-to-sentence” edge. This edge type simply joins an end-of-sentence punctuation node with the first word of the subsequent sentence. The intuition for this edge is that an entity that is mentioned in one sentence can be in a semantic relation with an entity in the adjacent sentence and that the likelihood of such a relation diminishes with increasing number of sentences that exists between the two entities.

The text graph in Figure 3.contains two sentence-to-sentence edges: one between the period in the first sentence and the first word (“The”) in the second sentence; the other between the period in the second sentence and the first word (“TcpC”) in the third sentence.

7.1.3 Coreference Nodes and Edges

Another source of intersentential edges that will be considered is coreference edges. These edges assume that in-document coreference resolution has been accurately performed. The intuition for this edge is that because the entities refer to the same thing, anything that is said in one sentence can apply to the entity in the next mentioning of the entity. We create a node for each entity and associate an edge between the node and each instance of the entity.

The text graph in Figure 3 contains three coreference edges. The edges all relate to the same entity “pilus” which we assume to be detected by a named-entity recognition system as be referring to the same concept.

7.2 Text-Graph Properties

This section describes some of the properties of the text graph defined above that will be exploited by the TeGRD algorithm.

1) The distance between two nodes is simply the number of edges between nodes.
2) For every pair of nodes *n* and *v* there is a walk from *n* to *v*. I.e. the graph is connected.
3) The graph can have cycles, and these cycles must involve coreference edges.

4) An entity instances E_i is in a *p-shortest path* relation with entity instance E_j if there are only p-1 other entity instances in a shorter shortest-path relation with E_i. The value of p can be interpreted as the rank of the proximity between the two entities, e.g. 1st nearest, 2nd nearest, etc. Assume that two entities that are tied in a *p-shortest path relation* are counted only once.

7.3 Feature Space

Given the above definition of a relation mention graph, each candidate relation mention is associated with a feature vector. The feature space will be identical to that proposed in (Melli & al, 2007) which was intended to subsume the feature space of the best performing method at the time particularly the proposal by (Jiang and Zhai, 2007).

Two types of features are proposed: global and local features. These relate in part to the “Structural” and “Content” features proposed in (Miller *et al*, 1998). Global features describe the overall structure of the graph. Local features describe in detail neighbourhoods within the graph. Table 12 illustrates the structure of the feature space with respect to each labelled relation mention.

Relation Case				Feature Space																				label
Passage	$E_{1,j}$	…	$E_{n,j'}$	Global Features										Local Features										
				Overall			Pair-wise							Overall			Pair-wise							
							$E_1 \leftrightarrow E_2$			…	$E_{n-1} \leftrightarrow E_n$						$E_1 \leftrightarrow E_2$			…	$E_{n-1} \leftrightarrow E_n$			

Table 12 – A tabular representation of the feature space along with the relation mention identifier and label assignment. Details of the features are presented below.

8: EMPIRICAL EVALUATION

8.1 Evaluation Overview

The proposed algorithm, `SDOI`, will be evaluated on the `smswo1` and `kdd09cma1` datasets, and compared against that several baseline algorithms on the performance measures described in section 2.3. Some preliminary results are available and included in this proposal.

8.2 Baseline Algorithms

We plan to compare `SDOI's` performance against baseline algorithms for each of the three subtasks.

For the mention identification task the baseline used selects the longest sequence of tokens that matches a concept's preferred name or synonym. This is the approach followed in (Kulkarni & al, 2009) and the non-Wikipedia experiments in (Milne & Witten, 2008). For our task, this baseline will likely achieve poor recall rate because it cannot identify concept mentions that are not yet in the ontology.

The main baseline algorithm for the linking task is based on the supervised approach proposed in (Milne & Witten, 2008). We reimplement the three features defined in their proposal (see Section 5: *RCI*, *AMW08rel*, and Σ*MW08*). We also reimplement the algorithm's two-phased approach to handle the two collective features used. The first phase selects a set of context concepts (S_{MW08}), and the second phase applies a binary classifier on the three features. We replicate the first phase by committing to all candidate concepts that pass tests t_1 and t_2, and that result in a single candidate concept. Separately, we also compare against the non-supervised approaches of either selecting a random concept (*RAND*$_B$), or selecting the concept with the most internal inlinks in the ontology (CI_B).

For the task of mention relation identification the benchmark algorithm will be to predict all pairs of linked concept mentions.

Finally, for the joint task of identification and linking both the baseline and the proposed `SDOI` algorithm simply apply their solutions to the two tasks in serial order. Each identified concept mention is passed to the linker.

8.3 Evaluation of Concept Mention Identification and Linking

This section will evaluate performance on the concept mention identification and linking tasks. Some of this work has been performed and included.

8.3.1 Candidacy Definition

Before proceeding to assessing `SDOI`'s performance, we first identify the subset of the eight candidacy test defined in Section 6.2 by empirical means. The definition of the candidacy selection heuristic can impact performance. Too restrictive a policy will limit the maximal attainable recall performance. Too liberal a policy could swamp the classifier with a large proportion of negative-to-positive training cases. We empirically test the effect on *F1* performance of incrementally adding individual tests (primary and stemmed[11]) in the following order: t_1+t_{1s}+t_2+t_{2s}+t_3+t_{3s}+t_4+t_{4s}. Table 13 summarizes the impact of sequentially adding each of the tests. Based on this empirical analysis the candidacy test select for SDOI included tests t_1+t_{1s}+t_2+t_{2s}+t_3+t_{3s}. Adding more tests beyond this point drops *F1* performance significantly, likely because the average number of training cases per mention increases from approximately 2.5 to 47 cases per mention on average.

Table 13 – Effect of the candidacy test definition on linking performance. As the test becomes more inclusive the maximum possible recall and number of training increases. The selected combination is highlighted.

Test	training cases	max. possible Recall	*SCMILO*		
			P	*R*	*F1*
+ t_1	536	9.0%	56.5%	7.8%	13.7%
+ t_{1s}	1,278	19.1%	61.3%	16.9%	26.4%
+ t_2	3,126	40.4%	62.7%	36.5%	46.1%
+ t_{2s}	5,206	53.0%	69.3%	34.4%	46.0%
+ t_3	9,390	74.8%	69.2%	48.0%	56.7%
+ t_{3s}	**11,598**	**77.3%**	**68.5%**	**49.2%**	**57.3%**
+ t_4	296,086	90.1%	72.8%	38.4%	50.3%
+ t_{4s}	386,537	91.5%	73.4%	37.1%	49.3%

To estimate algorithm performance we performed a leave-one-out cross-validation study. Specifically, we iterated through all 139 documents, leaving one document out of the training corpus and testing on all the mentions within the excluded document. The CRF++[12] package was used to generate the sequential tagging model used in the identification task. SVMlight[13] was used as the classification model training system used for the linking task. The number of iterations for the iterative classifier was set to five (μ=5).

Performance is reported on the separate tasks of: 1) predicted anchor text versus actual anchor text[14], 2) predicted concept node versus actual concept node in the ontology on the manually annotated concept mentions, and finally 3) predicted anchor text and concept node vs. actual anchor text and concept. Table 13 summarizes the performance of the SDOI and baseline algorithms.

Table 14 – Performance of the baseline algorithms and the proposed algorithm (Precision, Recall, F1-measure) on the two subtasks and the entire task.

Task		SDOI	Baseline
Identification	P	68.1%	73.1%

[11] http://search.cpan.org/perldoc?Lingua%3A%3AStem

[12] http://crfpp.sourceforge.net/

[13] http://svmlight.joachims.org/

[14] We used the evaluation script of the CoNLL-2000 chunking task http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

	R	71.4%	21.1%
	F1	69.7%	32.7%
LINKING	P	68.5%	55.4%
	R	49.2%	39.2%
	F1	57.3%	44.7%
IDENTIFICATION AND LINKING	P	42.7%	37.1%
	R	48.5%	11.6%
	F1	45.4%	17.7%

In the identification task the baseline attains a relatively high precision. The longest sequence match with ontology concept is likely to be a concept mention. However this rule came at significant expense in recall. The precision of SDOI's CRF model could be increased, but we opted to retain an optimal *F1*.

In the linking task SDOI performed better at linking concept mentions to the ontology than the baseline on all three measures. This is likely due to the additional features and the expanded definition of candidacy.

In the joint task the baseline performed poorly in terms of recall because of the cumulative effect of having performed poorly in the identification task. The baseline algorithm could not make a link prediction for the many mentions it failed to identify.

Separately Table 14 reports relative performance in the linking task of different features spaces against the two unsupervised approaches of random concept selection $RAND_B$, and most common selection CI_B. The table indicates a benefit of committing to all proposed features.

Table 15 – Relative performance lift in *F1* of different feature sets relative to the random $RAND_B$, and most common concept CI_B, baselines.

$RAND_B$	CI_B	FEATURE SET
-1.2%	-4.0%	Anchor Text-based only
3.1%	1.2%	Three features proposed in [9]
6.2%	3.5%	Anchor Text and Collective-based
17.7%	14.6%	All except Collective-based
18.2%	15.3%	All features

8.3.2 Collective Features & Iterative Classification

An unexpected result has been that the collective features contributed negligibly to overall linking performance. As seen in Table 14, excluding them from the feature space resulted in only a marginal reduction in linking performance. This is a surprising result given the lift attributed to collective features elsewhere in the literature.

We explored the possibility that the significantly expanded set of features that SDOI uses leaves fewer ambiguities that required deep insight into the roles of the concepts. As Table 15 shows, when only the anchor-text based features were retained the collective features begin to more noticeably contribute to the performance during each iteration.

Table 16 – Performance at each iteration on the full set of features, and on only a partial set of features. When fewer features are present the collective features are more relevant.

SCMILO	*F1* performance	
Iteration	All Features	Anchor-Text and Coll. Feat.
1	57.1%	47.5%
2	57.2%	48.4%
3	57.3%	49.4%
4	57.3%	49.9%
5	57.3%	50.2%

8.3.3 Mention-level Modeling

Some preliminary analysis of mention-level modelling has been performed. Table 17 summarizes the results. The four rows represent three different feature spaces (all features, minus collective feature, minus collective features and candidate set features). The three columns represent three alternative measures of performance: overall accuracy (correct/incorrect), linked mention accuracy (proportion of linked mentions that are correct), and proportion of accurately linked to inaccurately linked mentions

The results suggest that overall accuracy drops when the trained mention model is applied. One performance measure that is positively impacted is a type of precision metric based on the proportion of linked mentions with respect to the proportion of inaccurately linked mentions. This final measure recognizes a benefit for being more conservative about making link predictions.

If these results hold after further analysis, then this would be a negative result, which suggests that no additional modelling is necessary at the mention-level.

Table 17 - Performance comparison between the two different methods of selecting the final prediction: highest confidence rule or the trained mention model.

	accuracy					linked mention accurately linked?					link prediction (correct vs. incorrect)				
Feature Space	highest conf.	random cand.	commonness	mention mod. #1	mention mod. #2	highest conf.	random cand.	commonness	mention mod. #1	mention mod. #2	highest conf.	random cand.	commonness	mention mod. #1	mention mod. #2
All	63.0%	57.5%		62.6%	58.0%	41.4%	45.6%		38.9%	19.3%	65.5%	38.6%		66.6%	82.4%
All - coll.	62.4%	57.5%		62.1%	57.0%	40.1%	45.6%		38.8%	16.7%	63.6%	38.6%		64.4%	82.2%
All - cand set	56.8%	57.5%		56.6%	54.4%	19.6%	45.6%		18.5%	11.4%	64.6%	38.6%		65.0%	73.7%
All - coll. - cand set	56.3%	57.5%		56.1%	54.7%	17.7%	45.6%		17.3%	12.7%	64.2%	38.6%		64.1%	68.7%

8.4 Relation Mention Identification

This section will be based on the results of the application of TeGRR (as described in Section 7) on the evaluation data by also using a leave-one-document-out mechanism. The relation mentions within one document will be predicted based on a model trained on the remaining documents.

One technical challenge to this evaluation task is that a different version of the ontology must be created for each trained model because of the need to occlude any relation mentions that are only present in the document that is about to be tested.

8.5 Time Savings Evaluation

This section will be based on the results of a time savings evaluation. No such evaluation has been performed to date. The planned approach is to randomly pick twenty abstracts from the KDD-2008 conference. Half of the abstracts will be manually annotated from scratch; the other half will first by processed by SDOI and then manually fixed. Both annotation tasks will be timed and the resulting durations will be compared. For example if the manually annotated abstracts take two-hundred (200) minutes and the pre-annotated abstracts consumer eighty (80) minutes then the ratio will be 200/80 = 2.5. An interpretation of this ratio is that a human annotator can now annotated two and a half (2.5) documents in the same time that it previously took to annotate one (1) document.

9: CONCLUSION AND FUTURE WORK

We propose a thesis on the topic of concept and relation mention recognition with respect to an ontology by supervised means. Our main contribution will be the formalization of the task, the proposal of a state-of-the-art-algorithm (SDOI), and a publicly available benchmark dataset. The thesis will also cover related topics of the use of iterative classification and partial credit assignment, and may also present negative results such as the weak predictive of collective features when the feature space is enriched with local information.

The thesis will conclude with a discussion of possible future research directions. We foresee that this discussion will focus on ways to improve performance, but also to begin to apply SDOI in the real world. For example, an interesting direction for performance improvement is through a tight integration of the three subtasks so that, for example, information from the last task of relation mention identification can also be used to inform the first task of concept mention identification. Next, if performance results suggest that the solution can save annotation time then it would be interesting to expand the corpus to include all past KDD conference abstracts, and to expand the data mining ontology to include many of the main relations discovered in the process. Ideally we would like to integrate SDOI into the submission process of future conferences, such as KDD-2011, in order to have the authors themselves validate and correct the pre-annotated versions of their abstracts.

REFERENCE LIST

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. (2006). Link Prediction Using Supervised Learning. In: Proceedings of SDM-2006 Workshop on Link Analysis, Counter-terrorism and Security.

[2] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. (2001). Reducing Multiclass to Binary: a unifying approach for margin classifiers. In: The Journal of Machine Learning Research, 1. [doi>10.1162/15324430152733133]

[3] Satanjeev Banerjee, and Ted Pedersen. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: Proceedings of CICLing (2002). Lecture Notes in Computer Science; Vol. 2276.

[4] Indrajit Bhattacharya and Lise Getoor. (2004). Iterative Record Linkage for Cleaning and Integration. In Proceedings of KDD-2004.

[5] Mikhail Bilenko, Sugato Basu, and Mehran Sahami. (2005). Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM-2005).

[6] David M. Blei, and John Lafferty. (2006). A Correlated Topic Model of Science. In: Annals of Applied Statistics, 1(1). [doi>10.1214/07-AOAS114]

[7] Paul Buitelaar, Philipp Cimiano, Anette Frank, Matthias Hartung, and Stefania Racioppa. (2008). Ontology-based Information Extraction and Integration from Heterogeneous Data Sources. In: International Journal of Human-Computer Studies, 66(11).

[8] Rudi L. Cilibrasi, and Paul M. Vitanyi. (2007). The Google Similarity Distance. In: IEEE Transactions on Knowledge and Data Engineering 19(3). [doi>10.1109/TKDE.2007.48]

[9] Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh Mohania. (2006). Efficiently linking text documents with relevant structured information. In Proceedings of the VLDB Conference (VLDB 2006)

[10] Eugene Charniak. (2000). A Maximum-Entropy-Inspired Parser. In: Proceedings of NAACL Conference (NAACL 2000).

[11] Nitesh Chawla, Nathalie Japkowicz, Aleksander Kolcz. (2004). Editorial: Special issue on learning from imbalanced data sets. In: ACM SIGKDD Explorations Newsletter, 6(1). [doi>10.1145/1007730.1007733]

[12] Jeremiah Crim, Ryan McDonald, and Fernando Pereira. (2005). Automatically Annotating Documents with Normalized Gene Lists. In: BMC Bioinformatics 2005, 6(Suppl 1):S13.

[13] Silviu Cucerzan. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In: Proceedings of EMNLP-CoNLL-2007.

[14] Dayne Freitag, and Andrew McCallum. (1999). Information Extraction with HMMs and Shrinkage. In: Proceedings of the AAAI 1999 Workshop on Machine Learning for Information Extraction.

[15] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. (2008). Open Information Extraction from the Web. In: Communications of the ACM, 51(12). [doi>10.1145/1409360.1409378]

[16] Oren Etzioni, Michele Banko, and Michael J. Cafarella. (2006). Machine Reading. In: Proceedings of the 21st AAAI Conference (AAAI 2006).

[17] Johannes Fürnkranz. (2002). Round Robin Classification. In: The Journal of Machine Learning Research, 2. [doi>10.1162/153244302320884605]

[18] Rob Hall, Charles Sutton, and Andrew McCallum. (2008). Unsupervised Deduplication Using Cross-field Dependencies. In Proceedings of SIGKDD Conference (KDD 2008).

[19] Lise Getoor, and Christopher P. Diehl. (2005). Link Mining: A survey. In: SIGKDD Explorations, 7(2).

[20] Joseph Hassell, Boanerges Aleman-Meza, and I. Budak Arpinar. (2006). Ontology-driven automatic entity disambiguation in unstructured text. In: Proceedings of the 5th International Semantic Web Conference (ISWC).

[21] Jing Jiang, and ChengXiang Zhai. (2007). *A Systematic Exploration of the Feature Space for Relation Extraction.* In: Proceedings of NAACL/HLT Conference (NAACL/HLT 2007).

[22] Valentin Jijkoun, Mahboob Alam Khalid, Maarten Marx, and Maarten de Rijke. (2008). Named Entity Normalization in User Generated Content. In: Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data (AND 2008:23-30).

[23] Dmitri V. Kalashniko, Sharad Mehrotra, and Zhaoqi Chen. Exploiting Relationships for Domain-Independent Data Cleaning. In: Proceedings of the SIAM International Conference on Data Mining (SDM 2005)

[24] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. (2009). Collective Annotation of Wikipedia Entities in Web Text. In: Proceedings of ACM SIGKDD Conference (KDD 2009). [doi>10.1145/1557019.1557073]

[25] Michael Lesk. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from a ice cream cone. In: Proceedings of SIGDOC-1986.

[26] Andrew McCallum, and Wei Li. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: Proceedings of Conference on Natural Language Learning (CoNLL 2003).

[27] Bernardo Magnini, Emanuele Pianta, Octavian Popescu and Manuela Speranza. (2006). Ontology Population from Textual Mentions: Task Definition and Benchmark. In: Proceedings of the ACL 2006 Workshop on Ontology Population and Learning.

[28] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. (2008). Introduction to Information Retrieval. Cambridge University Press. ISBN 0521865719

[29] Gabor Melli, and Martin Ester. (2010a) Supervised Identification of Concept Mentions and their Linking to an Ontology. *Submitted*.

[30] Gabor Melli, and Martin Ester. (2010b) Evaluating and Training on Partially Correct Concept Mention Links. *Submitted*.

[31] Gabor Melli. (2010). Annotated Term Mentions in the KDD-2009 Abstracts Linked to a Data Analysis Ontology. In: Proceedings of 7th Int. Conference on Language Resources and Evaluation (LREC 2010) *forthcoming*

[32] Gabor Melli, and Jerre McQuinn. (2008). Requirements Specification Using Fact-Oriented Modeling: A Case Study and Generalization. In: Proceedings of Workshop on Object-Role Modeling (ORM 2008)

[33] Gabor Melli, Martin Ester, and Anoop Sarkar. (2007). Recognition of Multi-sentence n-ary Subcellular Localization Mentions in Biomedical Abstracts. In: Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM 2007).

[34] Gabor Melli, Yang Wang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar and Fred Popowich. (2005). Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task. In: Proceedings of Document Understanding Workshop at the HLT/EMNLP Annual Meeting (DUC 2005).

[35] Rada Mihalcea, and Andras Csomai. (2007). Wikify!: Linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM 2007). [doi>10.1145/1321440.1321475]

[36] David N. Milne, and Ian H. Witten. (2008). Learning to Link with Wikipedia. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, (CIKM 2008). [doi>10.1145/1458082.1458150]

[37] Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M. Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William W Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. (2008). Overview of BioCreative II gene normalization. In: Genome Biology 2008, 9(Suppl 2):S3. [doi>10.1186/gb-2008-9-s2-s3].

[38] Roberto Navigli, Paola Velardi, and Aldo Gangemi. (2003). Ontology Learning and Its Application to Automated Terminology Translation. In: IEEE Int. Systems, 18(1).

[39] Jennifer Neville, and David Jensen. (2000). Iterative Classification in Relational Data. In: Proceedings of the Workshop on Statistical Relational Learning.

[40] Ryan Rifkin, and Aldebaro Klautau. (2004). In Defense of One-Vs-All Classification. In: The Journal of Machine Learning Research, 5.

[41] Sunita Sarawagi. (2008). Information extraction. FnT Databases, 1(3), 2008.

[42] Francesco Sclano, and Paola Velardi. (2007). TermExtractor: A web application to learn the common terminology of interest groups and research communities. In: Proceedings of the 9th Conference on Terminology and AI (TIA 2007).

[43] Fei Sha, and Fernando Pereira. (2003). Shallow Parsing with Conditional Random Fields. In: Proceedings of Conference of the North American Chapter of the

Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003). [doi>10.3115/1073445.1073473]

[44] Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. (2006). Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006). [doi>10.1145/1150402.1150492]

[45] Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh Mohania. (2006). Efficiently Linking Text Documents with Relevant Structured Information. In: Proceedings of VLDB Conference (VLDB 2006).

[46] Gang Wu, and Edward Y. Chang. (2004). Aligning Boundary in Kernel Space for Learning Imbalanced Dataset. In: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM 2004) [doi>10.1109/ICDM.2004.10106]

[47] Min Zhang, Jie Zhang, and Jian Su. (2006). Exploring Syntactic Features for Relation Extraction using a Convolution Tree Kernel. In: Proceedings of HLT Conference (HLT 2006).

[48] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. (2003). Kernel Methods for Relation Extraction. In: Journal of Machine Learning Research, 3.

[49] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. (2007). Frontiers of Biomedical Text Mining: current progress. In: Briefings in Bioinformatics 2007, 8(5). Oxford Univ Press. [doi>10.1093/bib/bbm045]