Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports

Jiyin He
Centrum Wiskunde en Informatica
Science Park 123, 1098XG
Amsterdam, the Netherlands
J.He@cwi.nl

Maarten de Rijke
University of Amsterdam
Science Park 904, 1098XH
Amsterdam, the Netherlands
derijke@uva.nl

Merlijn Sevenster
Philips Research
High Tech Campus 34,5656AA
Eindhoven, the Netherlands
merlijn.sevenster@philips.com

Rob van Ommering
Philips Research
High Tech Campus 34,5656AA
Eindhoven, the Netherlands
rob.van.ommering@philips.com

Yuechen Qian
Philips Research
345 Scarborough Road
Briarcliff Manor, NY 10510, USA
yuechen.qian@philips.com

ABSTRACT

Automatically annotating texts with background information has recently received much attention. We conduct a case study in automatically generating links from narrative radiology reports to Wikipedia. Such links help users understand the medical terminology and thereby increase the value of the reports.

Direct applications of existing automatic link generation systems trained on Wikipedia to our radiology data do not yield satisfactory results. Our analysis reveals that medical phrases are often syntactically regular but semantically complicated, e.g., containing multiple concepts or concepts with multiple modifiers. The latter property is the main reason for the failure of existing systems. Based on this observation, we propose an automatic link generation approach that takes into account these properties. We use a sequential labeling approach with syntactic features for anchor text identification in order to exploit syntactic regularities in medical terminology. We combine this with a sub-anchor based approach to target finding, which is aimed at coping with the complex semantic structure of medical phrases. Empirical results show that the proposed system effectively improves the performance over existing systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems

General Terms

Experimentation, Algorithms

Keywords

Automatic link generation, Wikipedia, radiology reports

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

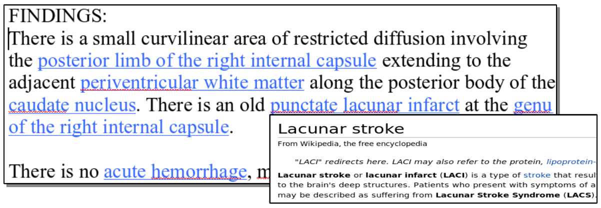
1. INTRODUCTION

Hypertext links are useful because they offer users a way of getting to pertinent information when they are not aware of that information or when they simply do not know enough about the topic at hand to be able to explicitly ask for it. Many types of links exist, e.g., categorical links (such as the navigation of a website), links to related items (such as linking events in news along a timeline), links to "similar items" (in book reviews, or in online shopping environments), etc. In this paper we focus on explanatory links; that is, the target of a link provides definitions or background information for the source of the link. We study the problem of automatically generating this type of link with Wikipedia: Given a piece of text, identify terms or phrases whose meaning is important for understanding the text and link them to Wikipedia pages for explanations or background information.

Automatic link generation with Wikipedia has received much attention in recent years [9, 20, 28, 29, 31]. Most of the studies are interested in solving a generic problem (e.g., developing an automatic link generation approach using Wikipedia as training material [29, 31]), or applying link generation techniques in general domains that cover diverse topics, (e.g., news, blogs, web, etc. [5, 9, 20, 28]). We consider a scenario in the radiology domain where we aim to link medical phrases found in radiology reports, which are typically anatomy or diagnosis terms, to corresponding Wikipedia concepts.

A radiology report gives a narrative account of the radiologist's findings, diagnoses and recommendations for followup actions. Radiology reports are the principal means of communication between radiologists and referring clinicians such as surgeons and general practitioners. The structure of radiology reports may vary over institutes but generally consists of several sections, including patient history, image protocols, findings and conclusions. Depending on the complexity of the cases, reports may have varied lengths, e.g., more than 40 lines of text for a complicated case.

Linking medical phrases to Wikipedia is a typical application of generating explanatory links and is useful in various realistic scenarios. For example, while reading a medical report, the patients concerned are usually not familiar with its medical terminology. By automatically identifying medical terms and explaining them through a link to a knowledge resource that is accessible and understandable by non-experts, the reports become more valuable for non-experts, thereby improving expert-patient communication. In Figure 1 we show an excerpt from a radiology report, with medical terms requiring explanation highlighted, together with a snippet from a Wikipedia page describing one of the highlighted medical terms.



FINDINGS:
There is a small curvilinear area of restricted diffusion involving the posterior limb of the right internal capsule extending to the adjacent periventricular white matter along the posterior body of the caudate nucleus. There is an old punctate lacunar infarct at the genu of the right internal capsule.

There is no acute hemorrhage, m

Lacunar stroke

From Wikipedia, the free encyclopedia

"LACI" redirects here. LACI may also refer to the protein, lipoprotein-

Lacunar stroke or **lacunar infarct** (**LACI**) is a type of stroke that resul to the brain's deep structures. Patients who present with symptoms of a may be described as suffering from **Lacunar Stroke Syndrome** (**LACS**).

Figure 1: An example of linking medical phrases in radiology report to Wikipedia.

There exist many medical knowledge source; Wikipedia turns out to be a competitive resource for the purpose of providing explanatory background material to laymen, for the following reasons. (i) Quantity: Wikipedia densely covers the medical domain; it contains medical lemmas from multiple medical thesauri and ontologies, such as International Statistical Classification of Diseases and Related Health Problems (ICD-9, ICD-10), Medical Subject Headings (MeSH), and Gray's Anatomy, etc. (ii) Quality: although written collaboratively by Internet volunteers, the quality of Wikipedia articles is guaranteed by the Wikipedia content criterion "verifiability," that is, material in a Wikipedia page should be verifiable against a reliable source; errors in the content are often spotted quickly and corrected by collaborative editors [39]. (iii) Accessibility: Wikipedia is a free online resource. All users can access its content without registering or creating an account. (iv) Readability: the content of Wikipedia is usually written at non-expert level.

Given the above scenario and motivation, we seek answers to the following research questions through empirical method:

RQ1 Can state-of-the-art automatic link generation systems, which are in principle domain independent, be effectively applied to annotating radiology reports? If not, how do links in radiology reports differ from links in a general domain, e.g., Wikipedia links?

RQ2 How can we exploit these differences to effectively annotate radiology reports with links to Wikipedia, and how does the resulting approach compare to domain-independent link generation systems that are state of the art in the field?

We evaluate two state-of-the-art systems, the Wikify! system [29] and the Wikipedia Miner system [31], on a test collection of radiology reports that have been manually annotated by domain experts with links to Wikipedia concepts; see Section 3 for details about the test collection. Neither system yields satisfactory results on our test collection. Two properties of medical phrases are key. First, they are often syntactically regular: they are mostly noun phrases with one or more modifiers (e.g., adjectives). Second, while syntactically regular, medical phrases often have a complicated semantic structure, due to the presence of multiple modifiers as well as conjunctions of concepts within a single phrase. The latter property is the major reason for the failure of both systems, as Wikipedia concepts are often relatively simple, e.g., consist of a single concept or concepts without modifiers. We propose an automatic link generation approach that takes these properties into account. We use a sequential labeling approach with syntactic features to identify anchor texts, i.e., medical phrases, in order to exploit syntactic regularities in medical terminology. We then combine this with a sub-anchor based approach to target finding, which is aimed at coping with the complex semantic structure of medical phrases. The proposed system is shown to effectively improve over the two state-of-the-art systems.

Our contribution is two-fold. First, our study shows that automatically generating links to Wikipedia for reports in the radiology domain requires non-trivial adaptations to the existing domain-independent linking systems; our analyses contribute to the understanding of the state-of-the-art linking systems as well as the problems that are specific to the radiology domain. Second, we propose a link generation approach that takes into account these domain-specific properties, which is shown to be able to effectively link medical concepts from radiology reports to Wikipedia concepts.

Section 2 below discusses related work. In Section 3, we introduce our notation and evaluation setup, including a test collection developed for evaluating automatic link generation for radiology data. In Section 4 we evaluate and analyze the performance of the two state-of-the-art link generation systems on this test collection. Based on our analysis, we introduce our proposed approach in Section 5. In Section 6 we evaluate it and include a comparison to the two state-of-the-art systems, followed by a further analysis and discussion in Section 7. Section 8 concludes the paper.

2. RELATED WORK

Automatic link generation with Wikipedia. Automatically generating hypertext links has a long history, going back nearly two decades. Early publications include [1, 3, 10, 14]. Later commercial approaches have met with limited success [32]. In the context of Wikipedia, renewed interest in automatic link generation emerged. A relatively early paper on the topic is [11], where the problem of discovering missing hypertext links in Wikipedia is addressed. The method proposed consists of two steps: first, clustering highly similar pages around a given page, and then identifying candidate links from those similar pages that might be missing on the given page. The main innovation is in the algorithm that is used for identifying similar pages and not so much in the link detection. Meanwhile, the task of disambiguating links to Wikipedia has received special attention as part of semantically oriented tasks such as named entity normalization. Cucerzan [9] uses automatically generated links to Wikipedia to disambiguate named entities in news corpora. Generalizing Cucerzan [9]'s work to user generated content with additional heuristics, Jijkoun et al. [20] focus on the named entity normalization task on blogs and comments. Recently, Meij et al. [28] study the problem in the scenario of semantic query suggestions, where each query is linked to a list of concepts from DBpedia, ranked by their relevance to the query.

In 2007, INEX, the INitiative for the Evaluation of XML retrieval, launched the Link-the-Wiki (LTW) track; the goal is to automatically generate links among Wikipedia pages as well as between Wikipedia and other another encyclopedia collection [13, 19]. Within the LTW track, various heuristics as well as retrieval-based methods have been proposed [13, 19]. One important issue discovered is that there exist many trivial links in Wikipedia, such as year, country, etc., which are actively rejected by human assessors [18]. Our case study is partly motivated by this finding, as the disagreement between human assessors and the existing Wikipedia links can be amplified when we switch to a specific domain such as the radiology domain where the language usage is expected to be very different from that in a general domain.

The work that is closest to ours has been presented in [29] and [31]. The Wikify! system [29] identifies the anchor texts from a given text based on term statistics derived from Wikipedia links. Further, the authors experimented with both knowledge-based and machine learning-based approaches for identifying the corresponding Wikipedia concepts for recognized anchor texts. Milne and Witten [31] tackled the problem with machine learning techniques; contextual information in the source text was used to determine the best related Wikipedia concepts, which in turn also served as features for anchor text detection. Their approach greatly improved the performance in terms of precision and recall over [29]. We will discuss further details of the two systems in Section 4 and analyze their performance on the task of generating links from radiology reports to Wikipedia concepts.

Information extraction and encoding for narrative radiology reports. A number of information extraction systems have been developed that focus on narrative medical reports of the kind that we are interested in. The Special Purpose Radiology Understanding System (SPRUS) [16] is an early system that extracts and encodes findings and interpretations from chest radiology reports. The authors experiment with syntactic extensions of SPRUS, reporting a 81% recognition rate in small scale experiments (10 reports) [17]. The MetaMap tool [2] extracts medical phrases from free text and maps them to the vocabulary of the Unified Medical Language System (UMLS) ontology. The Medical Language Extraction and Encoding System (MedLEE) [12] is a rule-based natural language processing system that extracts clinical information from radiology reports and encodes it using a controlled vocabulary. It reports 70% recall and 87% precision scores on identifying four diseases from a set of 230 radiology reports.

The automatic link generation task bears considerable similarity to automatically extracting concepts from free text: both revolve around recognition of medical phrases and mapping the phrases found to a controlled space. However, in the case of automatic link generation the elements in the controlled space are complex and lengthy (e.g., the contents of a Wikipedia page), whereas they tend to be simple and short in the case of ontologies (e.g., the synonymous labels of an ontology concept).

Unlike the systems mentioned above, our approach will be strictly based on machine learning techniques. While we choose Wikipedia as our target collection for the reasons mentioned in Section 1, our approach is sufficiently flexible to allow for extensions with handcrafted rules or expert knowledge in the form of features as well as to be adapted for generating links to other collections.

3. PRELIMINARIES

In this section, we introduce our notation and describe the experimental setup that we use to answer the two research questions introduced in Section 1.

Terminology and notation. Let $T = \{t_i\}_i$ be the set of *source texts*, e.g., our corpus of radiology reports. Let $W = \{d_i\}_i$ be the set of Wikipedia concepts, where each concept is represented by a Wikipedia page. We split the link generation task in three subtasks:

1. *Anchor detection* (AD): Given a source text t with ngrams $NG^t = \{ng_i\}_i$, extract a set of *anchors* $A^t \subseteq NG^t$, that is, phrases that we want to link to a Wikipedia page.
2. *Target candidate identification* (TCI): Given an anchor $a \in A^t$, select a set C^a of candidate *targets*, that is, the set that contains the Wikipedia page to which a is linked. The set C^a can be the entire Wikipedia collection W, but it is often more efficient to consider a subset of W.
3. *Target detection* (TD): Given a set of candidates C^a, select a target $w \in C^a$ to which a linked.

A link generation system maps each source text t to a set of L pairs (a, w), i.e., the links it establishes between anchors a in t and Wikipedia pages w. In the literature, the notion of *target finding* is used to refer to the combined TCI and TD task; for ease of exposition we split the target finding task in two.

We use the functions $AD(\cdot)$, $TCI(\cdot)$ and $TD(\cdot)$ to refer to the above components. For instance, AD is a function that maps t to a subset of NG^t, and TD maps the set $\{C^a : a \in AD(t)\}$ to $\{(a, w) : a \in AD(t), w \in C^a\}$.

Collection of radiology reports. Our collection of radiology reports consists of 860 anonymized neuroradiology reports, obtained from a US-based radiology institute, written in the 2001–2008 time frame. The reports were obtained by querying the institute's picture archiving and communication system for neoplasm cases in the optic nerve region.

The collection was annotated by three annotators, all with a background in medical informatics. The anchor texts are phrases referring to body locations, findings and diagnoses. In total, 29,256 links, i.e., anchor–Wikipedia page pairs, are extracted from the 860 reports, which can be resolved to 6,440 unique links. The set was created as follows.

The corpus was divided in three subsets; each subset was assigned to an annotator. Each annotator manually selected anchor texts in all reports assigned to him. For each anchor, the annotators use Wikipedia's search engine to search for the most appropriate page. If an anchor text does not have a directly matching Wikipedia page, a more general concept that reasonably covers the topic was sought. If no such page was found, no target was assigned. Thus every anchor was assigned at most one Wikipedia page.

The annotations of the three annotators were merged by concatenating their lists of anchor–target pairs. In case two annotators assigned different target pages to the same anchor (which happened in less than 5% of the anchors), disagreements were discussed among annotators and one annotator manually corrected this by picking the best target page. Then, the selections in the reports were updated on the basis of the concatenated list of anchor–target page pairs ensuring that two occurrences of the same phrase, possibly in different reports, are assigned the same Wikipedia page, if any. Note that this indicates that our anchor texts are not ambiguous within the test collection. This is a strong assumption in a general domain as well as in the medical domain. For instance, in the medical domain, "ventricle" is ambiguous as it may refer to a space in the heart as well as an area in the brain. In our corpus, however, it turned out to be a weak assumption as all reports are on the topic of neuroradiology and no ambiguous phrases were encountered during the annotation process.

Wikipedia collection. For annotation, the online version of Wikipedia during 2008 was used. For system development, we use the INEX 2009 Wikipedia collection [36] as our target collection. It is a snapshot of the October 8, 2008 dump of the English Wikipedia articles, with a total of 2,666,190 articles.

Note that the content of the Wikipedia collection may change over time. In order to resolve possible differences between different versions of Wikipedia, we map all the resulting urls to the online version. This is the same for both the state-of-the-art systems discussed in Section 4 and our proposed approach discussed in Section 5. Around 8% of the target urls in the annotation have a redirection at the time we evaluated the systems.

Evaluation setup. We evaluate an automatic link generation systems' performance in terms of precision, recall and F-measure. We evaluate the systems' performance on each radiology report, and show their overall performance, i.e., the averaged performance over all reports. Further, we use a paired t-test for significance testing. The symbol ▲ (▼) indicates a significant increase (decrease) with p-value < 0.01; and a △ (▽) indicates a significant increase (decrease) with p-value < 0.05.

4. A SOLVED PROBLEM?

In order to answer the first research question, RQ1, we evaluate and analyze the performance of two state-of-the-art systems, i.e., Wikify! [29] and Wikipedia Miner [31] in automatically generating links from radiology reports to explanatory Wikipedia concepts. We start with a brief introduction of the two systems and specify our implementation.

4.1 Two state-of-the-art systems

Wikify!. The workflow of Wikify! [29] is summarized in Algorithm 1 using the notations introduced in Section 3. For detecting anchor texts from NG^t, i.e., $AD(\cdot)$, Wikify! ranks $ng \in NG^t$ according to a score and uses the top τ ranked ng's as anchor texts for t. Mihalcea and Csomai [29] experimented with several scores, including tf.idf, χ^2 and a *keyphraseness* score, which turns out to be the most effective score among the three. The keyphraseness score is defined as follows.

Algorithm 1 Wikify!

```
Input: $t$
$L = \emptyset$
for $a \in AD(t)$ do
    $C^a = TCI(a, W)$
    $w = TD(a, C^a)$
    $L \leftarrow L \cup \{(a, w)\}$
end for
return $L$
```

$$\textit{keyphraseness}(ng) = \frac{|A_{ng}|}{|W_{ng}|}, \tag{1}$$

where $|W_{ng}|$ is the number of Wikipedia pages that mention the ngram ng and $|A_{ng}|$ is the number of Wikipedia pages where ng occurs as anchor text.

Wikify! collects C^a for a given anchor a via existing links in Wikipedia. Whenever a is used as an anchor text in Wikipedia, the page it links to is added to the set of candidate targets C^a.

To identify the target page w from C^a for a given a, two approaches were proposed [29]. The first one is knowledge based and picks out the candidate target page w that maximizes the score calculated by the Lesk algorithm [24], i.e., the word overlap between the candidate page w and the context of a in t. The second is a machine learning-based approach. For each a, a classifier is trained to classify whether it links to a candidate target page.

We re-implemented the Wikify! system as described in [29]. For anchor detection, following [29], we set the threshold τ to 6% of the length of the source text. For target finding, we implement both the knowledge based approach and the machine learning based approach. A combination of the two approaches was evaluated in [29], but no consistent improvement was found in terms of precision, recall and F-measure, therefore we decide to leave it out.

Wikipedia Miner. The workflow of Wikipedia Miner [31] is summarized in Algorithm 2. Unlike Wikify!, Wikipedia Miner performs $TCI(\cdot)$ and $TD(\cdot)$ on ngrams instead of identified anchors.

Algorithm 2 Wikipedia Miner

```
Input: $t$
$L = \emptyset$, $L_{tmp} = \emptyset$
for $ng \in NG^t$ do
    $C^{ng} = TCI(ng, W)$
    $w = TD(C^{ng}, ng)$
    $L_{tmp} \leftarrow L_{tmp} \cup \{(ng, w)\}$
end for
for $(ng, w) \in L_{tmp}$ do
    if $ng \in AD(t)$ then
        $L \leftarrow L \cup \{(ng, w)\}$
    end if
end for
return $L$
```

To collect target candidates C^{ng}, Wikipedia Miner uses existing Wikipedia links. To improve efficiency, a threshold is used to filter out candidate pages that have very low chance of being linked to a given ng based on inspecting existing Wikipedia links. Wikipedia Miner trains a classifier for target detection. For details of the features as well as the combination of features employed in the system, we refer to [31]. Wikipedia Miner does not have an explicit anchor detection phase, instead, anchor detection is achieved by filtering the set of pairs from L_{tmp}. A classifier is trained over instances consisting of ngram-target pairs. Various features are used to train the classifier, including the keyphraseness score proposed in [29] and features reflecting the relatedness between source text and candidate target page.

For evaluation, we use the online Wikipedia Miner server,[1] which is provided by the authors, with default parameter settings. The server was accessed remotely and used as a black box.

4.2 Evaluation of the two systems

Results. Table 1 shows the evaluation results in three aspects: anchor detection, target finding and overall system performance. Here, target finding is evaluated over the anchor texts that are correctly identified by each system.

Table 1: Results of the two systems, Wikify! and Wikipedia Miner(WM) on the test collection of radiology reports, in terms of precision (P), recall (R) and F-measure (F).

System	Anchor detection			Target finding			Overall		
	P	R	F	P	R	F	P	R	F
Wikify! (Lesk)	0.35	0.16	0.22	0.40	0.40	0.40	0.14	0.07	0.09
Wikify! (ML)	0.35	0.16	0.22	0.69	0.69	0.69	0.25	0.12	0.16
WM	0.35	0.36	0.36	0.84	0.84	0.84	0.29	0.30	0.30

We see that for both systems, the overall performance is far from their performance on the data from general domain as reported in [29] and [31].[2] Further, both systems show a relatively reasonable performance of target finding on correctly identified anchor texts compared to the reported scores achieved in the general domain. This observation indicates that the major bottleneck of the systems' performance occurs in anchor detection. Given this observation, below we conduct an analysis of the anchor texts aiming at finding the answer to the question: how do links in radiology

[1] http://wikipedia-miner.sourceforge.net

[2] Milne and Witten [31] report a precision/recall of 0.73/0.74 on overall system performance; Mihalcea and Csomai [29] do not report overall system performance, but an upper bound is estimated to be 0.53/0.55 in terms of precision/recall [31].

Table 2: The number of annotated anchor texts/sub-anchors in radiology reports covered by Wikipedia anchor texts.

Evaluation type	Occur. in WP links	coverage (%)
exact match	923	14.3
partial match	1,038	16.1
sub-exact match	5,257	81.6

reports differ from links in a general domain, i.e., Wikipedia links? What causes the failure of the state-of-the-art systems?

An analysis of radiology anchor texts. Upon closer inspection of the anchors in the test collection, we have the following observations. Medical phrases often have a regular syntactic structure. For example, 70% of our annotated anchor texts are noun phrases, where 38% are single-word nouns and 32% are nouns with one or more modifiers (e.g., adjectives). Such regularity can provide useful features for recognizing these medical phrases in the reports.

Furthermore, Wikipedia anchors are generally shorter and less complex than radiology anchors. The average length of the annotated anchor texts in radiology reports is 3.29 words, while the average length of the title of their corresponding targets in Wikipedia is 1.98 words. Often the presence of multiple modifiers as well as conjunctions within a single medical phrase results in a complex semantic structure. For example, the phrase "acute cerebral and cerebellar infarction" contains two concepts "cerebellar infarction" and "acute cerebral infarction," where "cerebellar" and "cerebral" are synonyms. When linking this phrase to Wikipedia, one needs to identify the main concept it represents prior to searching for a target page in Wikipedia. Thus, there is a structural mismatch between the Wikipedia anchors and the anchors in the radiology reports.

In order to quantify the above mentioned difference between radiology anchor texts and Wikipedia anchors, in Table 2 we list a set of statistics about the coverage of Wikipedia anchor texts over the annotated anchor texts found in our test collection. Let A^W be all the anchor texts found in Wikipedia. We evaluate the coverage on three aspects:

exact match the number of annotated anchor texts occurring in A^W;

partial match the number of annotated anchor texts occurring in A^W, including the cases when an annotated anchor text is a substring of a Wikipedia anchor; e.g., the phrase "arachnoid" partially matches the Wikipedia anchor "arachnoid mater".

sub-exact match the number of annotated anchor texts containing at least one sub word sequence that occurs in A^W. For example, "left internal capsule" sub-exact matches the Wikipedia anchor "interal capsule".

Relatively few (<20%) annotated anchor texts occur (fully or as as a substring of the anchor texts) in A^W. However, over 80% of the annotated anchor texts do contain one or more sub-word sequences that occur in A^W. That is, most of the annotated anchor texts contain one or more Wikipedia concepts.

Now let us look at what these statistics mean for the system performance. For anchor detection, both state-of-the-art systems rely heavily on the Wikipedia anchor texts. The keyphraseness score is used as the only score for identifying anchor texts in the Wikify! system, and used as an important feature for Wikipedia Miner. However, from Eq. 1, we can see that an anchor text only receives a non-zero score if it occurs in A^W. Given the low coverage of the annotated anchor texts in A^W, it is not surprising that the keyphraseness score is not effective, as around 85% of the annotated anchor texts would receive a 0 score.

For target detection, both systems retrieve candidate target pages via Wikipedia links. The difference between the systems can be explained as follows. Recall that Wikify! finds its candidate target pages with respect to an identified anchor text, whereas Wikipedia Miner selects its candidate target pages from the set of all ngrams extracted from a report. The approach employed by Wikify! suffers from the same problem as in anchor detection: low coverage of Wikipedia anchor texts over annotated anchor texts in our test collection. In the case of Wikipedia Miner, since all possible ngrams in a report are considered, the whole pool of candidate target pages at the report level cover a majority of the annotated target pages for that report, which leads to a much improved performance compared to that of Wikify!.

In summary, the main reason that the state-of-the art systems do not yield a satisfactory performance in generating links from radiology data is caused by the structural mismatch between the radiology anchors and the Wikipedia anchors and the fact that both systems rely on existing Wikipedia links. Based on this observation, we now introduce our proposed approach aimed at tackling these problems.

5. LINK GENERATION REVISITED

How can we exploit the unique properties of radiology reports identified above? In order to exploit the regularity of the syntactic structure of medical phrases in the radiology reports, we treat the anchor detection problem as a sequential labeling problem. Sequential labeling is an effective approach in part-of-speech tagging as well as terminology recognition, including terminology from the biomedical domain [7, 21, 23, 27, 37, 38, 40]; it has been shown to be able to effectively capture the regularity in language usage. In addition, unlike the two state-of-the-art link generation systems, we learn the pattern of anchor texts from radiology data instead of Wikipedia. Wikipedia anchors are restricted to single concepts, while the radiology anchors can contain one or more concepts, or concepts with multiple modifiers, therefore Wikipedia anchors may not provide effective training material for sequential labeling to determine the boundaries of anchors in radiology reports.

Note that even though we expect that the sequential labeling approach can be useful in determining anchor text boundaries, the structural mismatch between Wikipedia anchors and radiology anchors remains for the target finding stage. To cope with the complex semantic structure of medical anchor texts, we propose a *sub-anchor-based* approach to retrieving candidate targets and to formulate features for target detection. By retrieving target candidates with respect to sub word sequence of an anchor text, referred to as *sub-anchors*, we collect candidate pages that are potentially relevant to different related concepts contained in the anchor text. During the target detection phase, we aggregate features extracted at sub-anchor level to anchor level. The feature of a single sub-anchor text is weighted by its relevance to the original anchor, as measured by its similarity to the original anchor text.

The proposed system is called LiRa, after Linking Radiology reports. Its workflow is the same as that of the Wikify! system, as illustrated in Algorithm 1. That is, it follows the following steps: anchor detection, target candidate identification and target detection. In the rest of this section, we first describe our approaches to the three components (Section 5.1-5.3), followed by the training and configuration details of the proposed system in Section 5.4.

5.1 Anchor detection

We define the sequential labeling task for anchor detection as follows. Given a text document, identify anchor texts by annotating each of the words in the text with one of the following labels: begin of anchor (BOA), in anchor (IA), end of anchor (EOA), outside anchor (OA), and single word anchor (SWA). SWA defines a single word anchor; BOA-(IA)-EOA defines an anchor with multiple words. Within this framework, we use a conditional random fields (CRF) model [22], which has shown state-of-the-art performance in solving sequential labeling problems [27, 37].

Let $WS = w_1, \ldots, w_n$ be an observed word sequence of length n, and $SS = s_1, \ldots, s_n$ a sequence of states where s_i corresponds to the label assigned to the word w_i. Following Settles [37], we use linear-chain CRFs, which define the conditional probability of the state sequence given the observed word sequence as

$$p(SS|WS) = \frac{1}{Z(WS)} \exp \sum_i^n \sum_k^m \lambda_k f_k(s_{i-1}, s_i, w_i, i), \quad (2)$$

where $Z(WS)$ is a normalization factor over all state sequences, $f_k(\cdot)$ is a feature function and λ_k is a learnt weight for feature $f_k(\cdot)$. The feature function describes a feature corresponding to the position i of the input sequence, states at position i and $i-1$, and word at position i.

The goal of the learning procedure is to find feature weights λ that maximize the log-likelihood of the training data:

$$LL = \sum_i \log p(s_i|w_i) - \sum_k \frac{\lambda_k^2}{2\sigma^2}. \quad (3)$$

The second term in Eq. 3 is a spherical Gaussian weight prior [6] used to penalize the log-likelihood term to avoid over-fitting.

In the literature various types of features have been explored, particularly syntactic features such as part-of-speech (POS) tags and orthographical features such as the combination of digits and letters [37, 40]. This is due to the fact that biomedical terminology, such as gene and protein names, often displays syntactic regularities as well as uncommon word spellings. Here, we expect that our medical phrases share the same properties as the biomedical terminology. Following the literature, we explored a number of features, including the word itself, its POS tag, its syntactic chunk tag, orthographical features as well as bigram and trigram features. Preliminary experiments show that three features, namely the word itself, its POS tag and syntactic chunk tag, are the most effective features. Adding other features does not result in significant improvement to the system's performance. Therefore in the rest of the paper we focus on these three basic features.

5.2 Target candidate identification

For an identified anchor text a, we retrieve Wikipedia pages related to the sub word sequences contained in a, i.e., sub-anchors, so as to collect candidate target pages with respect to different concepts related to the original anchor text.

Given a, we decompose it into the set of all sub-sequences $S_a = \{s_i\}_i$, while keeping the original order of the words within the identified anchor text. For example, for the anchor text "white matter disease", we have a set of sub-sequences {"white", "matter", "disease", "white matter", "matter disease", "white disease", "white matter disease"}.

In addition, there exist redirect pages in Wikipedia, which provide synonyms or morphological variations for a concept. For example, the concept "acoustic schwannoma" is redirected to "vestibular schwannoma." While decomposing an identified anchor text, we add its redirects to the set of sub-anchors. Further, in order to reduce the morphological variance of terms and phrases, we preprocess the radiology reports and the Wikipedia collection using the Porter stemmer [35]. In order to avoid adding trivial phrases, we filter out sequences that consist entirely of function words.

For each sub-anchor s, we retrieve a set of candidate target pages $C^s = \{c_i\}_i$, ranked in descending order of their *target probability*. Let $L_{s,c} = \{l(a, d)|a = s, d = c, d \in W\}$ denote all pairs of links found between s and c in Wikipedia links, that is, links between target page s and all occurrences of s as anchor texts. The target probability is calculated as

$$p(c_i|s) = \frac{|L_{s,c_i}|}{\sum_{j=1}^n |L_{s,c_j}|}, \quad (4)$$

where $n = |C^s|$ and $|\cdot|$ is the number of elements in a set. In other words, the target probability indicates how likely an anchor text is linked to a candidate page. An underlying assumption is that, the more an anchor text is linked with a candidate page in the existing Wikipedia links, the more likely it is that the anchor and the target page are closely related.

When examining the occurrence of sub-anchor s in existing Wikipedia links, we consider partial matches of phrases. That is, if all terms in s appear ordered within a Wikipedia anchor text, it is considered to be an occurrence.

We collect the top-K Wikipedia pages in terms of their target probability scores for each sub-anchor and use the union of all the collected pages from each sub-anchor as the candidate target pages for the anchor. While not all sub-anchor texts are semantically related to the original anchor texts, nor are they necessarily meaningful phrases, we leave the task of identifying (among all retrieved candidate target pages) the target page that is most relevant to the original anchor text, to the target detection component.

5.3 Target detection

We use a machine learning based approach to identify the target page d^* for a given anchor text a. Specifically, we train a classifier over the *anchor–candidate target* pairs (a, c), which are labeled as "link" or "non-link". We extract the following three type of features to train the classifier: (i) title matching, (ii) target probability, and (iii) language model log-likelihood ratio. The first two features are calculated at the sub-anchor level, and the third feature is calculated for a candidate target page; see below.

Title matching. One important feature of Wikipedia is that the title of each Wikipedia page represents the concept explained by the page. A match between an anchor text and the title of a Wikipedia page is therefore a strong indication that the content of the page is about the concept represented by the anchor text. The title matching score reflects the degree of matching between the anchor text and the title of c.

We consider the title matching scores for each of the sub-anchors. For a sub-anchor s of anchor a, and a candidate target page c, the title matching score is defined as follows:

$$tm(s, c) = f_{tm}(s, c)\frac{len(s)}{len(a)}, \quad (5)$$

where

$$f_{tm}(s, c) = \begin{cases} 1 & \text{if } s \text{ equals title of } c \\ 0 & \text{otherwise.} \end{cases}$$

and $len(\cdot)$ is number of words in a word sequence. The longer the sub-anchor, the more similar the sub-anchor is to the original anchor text, and therefore we have a higher degree of matching between the anchor text and the title of c. That is, a higher degree of word overlapping between the anchor text and the title of $ctar$.

Target probability. The target probability is the probability that a Wikipedia page will be selected as target page, given the anchor text, as defined in Eq. 4. It is pre-computed during the candidate retrieval procedure and calculated for each sub-anchor s.

Since the target probability is calculated at the sub-anchor level, we need to aggregate those scores for the original anchor texts.

Note that in the case of title matching, no explicit aggregation is needed, since for a given candidate target page, it can only match one of its sub-anchors. In the case of candidate target probability, we aggregate the features extracted from the sub-anchors into three features. For an anchor a and its sub-anchors S_a of a candidate target page c we define:

$$\max_{tp}(c) = \max_{s\in S} p(c|s);$$
$$\min_{tp}(c) = \min_{s\in S} p(c|s);$$
$$\text{wsum}_{tp}(c) = \textstyle\sum_{s\in S_a} \frac{len(s)}{len(a)} p(c|s).$$

Language-model log-likelihood ratio (LM-LLR). We use the LM-LLR as a feature to measure the extent to which a candidate target page is about radiology. It will help us to discriminate between e.g., the Wikipedia page about the journal *Brain* and the page about the body part that we are interested in.

Language models are statistical models that capture the statistical regularities in generating a language, often used to estimate the probability of a text being relevant to certain topic in the context of information retrieval [34]. Here we consider two language models. The first, θ_R, models the language used in the radiology reports, which we refer to as the *radiology model*, and the second, θ_W, models the language used in Wikipedia pages on topics in a general domain, which we refer to as *Wikipedia model*.

Each model defines a probability mechanism, which can be explained as follows. Assuming the two models sample terms from the radiology collection and the Wikipedia collection that follow a multinomial distribution, using a maximum likelihood estimation, the probability that a certain term t is selected given a collection C can be estimated as the relative frequency of the term in the collection, i.e., $p(t|\theta_C) = \frac{count(t\in C)}{|C|}$. Now, given a piece of text with n terms, $T = \{t_i\}_{i=1}^n$, the two models repeatedly sample n times, assuming independence between successive events. The probability that T is generated by the radiology model can be defined as

$$p(t_1, t_2, \ldots, t_n|\theta_R) = \textstyle\prod_{i=1}^n p(t_i|\theta_R), \tag{6}$$

while the probability that T is generated by the Wikipedia model is

$$p(t_1, t_2, \ldots, t_n|\theta_W) = \textstyle\prod_{i=1}^n p(t_i|\theta_W). \tag{7}$$

Given the above language models, we use the log-likelihood ratio (LLR) [26], a widely used model-comparison metric, to decide which model is more likely to have generated T:

$$\begin{aligned} \textit{LM-LLR}(T) &= \log\left(\frac{p(T|\theta_R)}{p(T|\theta_W)}\right) \\ &= \textstyle\sum_{i=1}^n \log p(t_i|\theta_R) - \sum_{i=1}^n \log p(t_i|\theta_W). \end{aligned} \tag{8}$$

To avoid zero probabilities, which come up if terms in T do not occur in the radiology reports or in Wikipedia, we use Laplacian smoothing [25]. That is, we assume that each word has been seen at least once. The LM-LLR score indicates which of the two models θ_R and θ_W is most likely to have generated T. A score larger than 0 indicates T is more likely to be generated by the radiology language model, hence more likely to be relevant to the anchor text identified from a radiology report.

In summary, we list the final features we use to train a classifier for identifying a target page from a set of candidate targets: (i) Title matching between a and c; (ii) Maximum target probability $\max_{tp}$; (iii) Minimum target probability $\min_{tp}$; (iv) Weighted sum of target probability wsum_{tp}; (v) Language model log-likelihood ratio of c.

5.4 Training and configuration

We specify the training procedure and the configuration of LiRa with respect to the three components described above.

Anchor detection. For anchor detection, we use the CRFsuite [33] implementation of CRFs with default parameter settings. A POS tagging and chunking tool TagChunk[3] is used to create the POS and chunk features. For training and evaluating the anchor detection performance, we use 3-fold cross-validation.

Target candidate identification. At the target candidate identification stage, we rank Wikipedia pages in descending order of target probability scores and select the top K candidate target pages. Heuristically, we set K to 10.

Target detection. We calculate the LM-LLR feature using the first 100 words of each candidate target page, for two reasons. First, the first paragraph of a Wikipedia page is usually the summary of the content of that page which covers the most important content of that page. Second, by using a constant number of words from each candidate target page, we eliminate the effect that the total number of words in the page has on the LM-LLR score. This makes the LM-LLR scores comparable across different Wikipedia pages.

Following [31], we experimented with three classifiers: Naive Bayes (NB), SVM [8], and Random Forest (RF) [4],[4] using the Weka implementations [15]. After preliminary experiments, we found that RF significantly outperforms the other two classifiers, in terms of efficiency and effectiveness. Therefore, all of LiRa's results reported below are based on RF. Further, with respect to the five features introduced in previous sections, we find in a preliminary feature selection experiment that all features are important to the classifier, removing any causes a decrease in performance.

Along with the predicted labels, the classifiers also provide a prediction confidence score. After classification, we execute a post-processing procedure. For anchor texts whose candidate target pages are all classified as "non-link," we select the candidate target with the lowest prediction confidence for being a "non-link." For anchor texts that have multiple candidate target pages labeled "link," we choose the one with the highest prediction confidence.

Again, we use 3-fold cross-validation for training and evaluating the classifiers.

6. EVALUATION

In this section, we show the evaluation results and compare the performance of the three link generation systems.[5]

6.1 Anchor detection

Table 3 lists the results of anchor detection for the three systems considered. We observe that LiRa outperforms both Wikipedia Miner and Wikify! in anchor detection in terms of all three evaluation metrics, i.e., precision, recall and F-measure. This suggests that the sequential labeling-CRF approach trained on radiology data is more effective than the approaches employed by Wikify! and Wikipedia Miner, which learn the patterns of anchor texts solely from existing Wikipedia links.

[3] http://www.umiacs.umd.edu/~hal/TagChunk/

[4] We use Random Forest, an ensemble decision tree classifier, instead of C4.5.

[5] In all the tables listed in this Section, boldface indicates the best performance across systems. For significance testing we use the same setting as described in Section 3. All runs are compared against LiRa.

Table 3: Results on anchor detection.

System	precision	recall	F-measure
LiRa	**0.90**	**0.80**	**0.85**
Wikipedia Miner	0.35▼	0.36▼	0.36▼
Wikify!	0.35▼	0.16▼	0.22▼

6.2 Target finding

In order to compare the target finding performance of the systems, we need to run the target finding components of each system on the same set of anchor texts. Consequently, we cannot simply evaluate each system on the set of anchors it finds in the anchor detection phase, as it is bound to differ from the set of anchors found by the other two systems. We consider two sets of anchor texts for evaluation. The first contains the annotated anchors from the test collection. However, since we run Wikipedia Miner as a black box, we cannot instruct it to generate a Wikipedia page for a given anchor. Therefore we only use this set to compare our system against Wikify!. The second set contains the anchor texts identified by Wikify! or by Wikipedia Miner. That is, we run LiRa on the anchor texts identified by Wikify! (Wikipedia Miner), and compare the target finding performance of LiRa against that of Wikify! (Wikipedia Miner, respectively) on the same set of anchor texts.

Table 4 shows the target finding performance of LiRa and Wikify! on the annotated anchors from the test collection. Table 5 (Table 6, respectively) shows the performance of LiRa and Wikify! (Wikipedia Miner, respectively) on the annotated anchor texts that are correctly identified by Wikify! (Wikipedia Miner).

In Table 4 and 5 we see that the target finding performance of LiRa is better than that of Wikify!.

Further, in Table 6 we see that LiRa also outperforms Wikipedia Miner on the target finding task, on the anchor texts found by the latter. The difference in performance between LiRa and Wikipedia Miner is smaller than the difference between LiRa and Wikify!.

Table 4: Comparing the performance of LiRa and Wikify! on target finding. The target finding algorithms are run on the annotated anchor texts found in the ground truth.

System	precision	recall	F-measure
LiRa	**0.68**	**0.68**	**0.68**
Wikify! (Lesk)	0.13▼	0.13▼	0.13▼
Wikify! (ML)	0.26▼	0.26▼	0.26▼

Table 5: Comparing the performance of LiRa and Wikify! on target finding. The target finding algorithms are run on the anchor texts identified by Wikify!.

System	precision	recall	F-measure
LiRa	**0.80**	**0.80**	**0.80**
Wikify! (Lesk)	0.40▼	0.40▼	0.40▼
Wikify! (ML)	0.69▼	0.69▼	0.69▼

Table 6: Comparing the performance of LiRa and Wikipedia Miner on target finding. The target finding algorithms are run on the anchor texts identified by Wikipedia Miner.

System	precision	recall	F-measure
LiRa	**0.89**	**0.89**	**0.89**
Wikipedia Miner	0.84▼	0.84▼	0.84▼

6.3 Overall performance

In Table 7 we show the overall performance of the three systems, i.e., the performance on the combined anchor detection and target finding task. We see that LiRa dramatically outperforms the state-of-the-art systems in terms of overall performance. This is no surprise as we have seen above that it also outperforms the other systems on the core subtasks, anchor detection and target finding.

Table 7: Overall system performance.

System	precision	recall	F-measure
LiRa(LL)	**0.65**	**0.58**	**0.61**
Wikipedia Miner	0.29▼	0.30▼	0.30▼
Wikify! (Lesk)	0.14▼	0.07▼	0.09▼
Wikfy! (ML)	0.25▼	0.12▼	0.16▼

Further, Table 4 can be seen as an "oracle run" of the overall system performance of Wikify! and LiRa, under the assumption that the systems' anchor detection modules are flawless, since we evaluate them on the anchors from the test collection as if the systems' found these anchors themselves. Again, we see that LiRa outperforms Wikify!. However, the performance of LiRa is far from perfect.

7. DISCUSSION AND ANALYSIS

In the previous section, we compared our approach to the state-of-the-art. LiRa consistently outperforms Wikify! and Wikipedia Miner on all aspects of the link generation task. Earlier findings in Section 4 help us understand the difference in performance.

For anchor detection, LiRa exploits the regularity of the syntactic structure of the annotated anchor texts in the radiology reports by using syntactic features within a sequential labeling approach. The sequential labeling approach captures this regularity, and is therefore more effective for anchor detection.

Further, even though all three systems retrieve candidate target pages on the basis of existing Wikipedia links, their performance differs considerably on target detection. Recall that Wikify! selects candidate target pages on the basis of identified anchor texts. It becomes clear that the marginal overlap between the Wikipedia and radiology anchors also hampers Wikify!'s target detection approach. The sub-anchor matches used by LiRa allow it to retrieve not only candidate pages that are relevant to the anchor text itself, but also pages that are relevant to related concepts, which improves the target detection performance. Recall that Wikipedia Miner uses yet another strategy to deal with the overlap between Wikipedia and radiology anchors: all possible ngrams in a radiology report that match Wikipedia anchor texts are considered, which results in a set of candidate target pages that cover the majority of the annotated target pages for that report. We saw that Wikipedia Miner's target finding performance is comparable to that of LiRa's.

Do linking systems perform differently with respect to different anchors? Are there anchors more difficult than others in terms of being correctly recognized and linked to a target page? While anchor texts can "differ" in many aspects, we decide to focus on one specific aspect: their frequency. In the test collection, we observe that some anchor texts occur more often than others. Specifically, in Figure 2, we rank the annotated anchor texts in decreasing order of their frequency and then plot their frequency against their ranks on a log scale. We see that the anchor text frequencies exhibit the typical properties of Zipf's law [30]. That is, a distribution consisting of a few anchors with high frequencies and a long tail of anchors with low frequencies. In addition, Table 8 lists the most frequent and least frequent anchors in our test collection. We see that

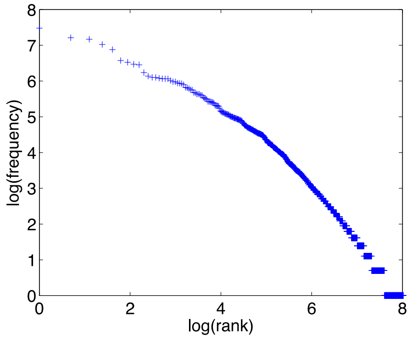


Figure 2: Distribution of anchor frequency. Anchors are ranked according to their frequency of occurrence in the radiology reports. The X-axis shows the logarithm of the ranks of anchors, and the Y-axis shows the logarithm of the frequency of the anchor at that rank.

frequent anchors tend to refer to an "easier" concept than the infrequent ones, i.e., with which (non-medical) users are more likely to be familiar. For instance, it is reasonable to assume that more users require background information about "xanthogranulomas" than on "brain" if such a term appears in a radiology report. Intuitively, frequent phrases are likely to refer to general concepts in the radiology reports while infrequent ones are likely to be used to describe a specific medical condition. Further, we expect that general concepts are more likely to occur in Wikipedia and therefore link generation is likely to be relatively easier for frequent phrases than for infrequent ones.

Table 8: The five most frequent and five least frequent anchor texts in the test collection.

Top 5	Bottom 5
mass	vestibular nerves
brain	Virchow-Robin space
meningioma	Warthin's tumor
frontal	Wegner's granulomatosis
white matter	xanthogranulomas

Based on the above observations and assumptions, we conduct a further analysis aimed at finding out whether and how anchor text frequencies have an impact on the performance of linking systems.

Table 9: Segmentation of anchor texts based on their frequencies in the test collection.

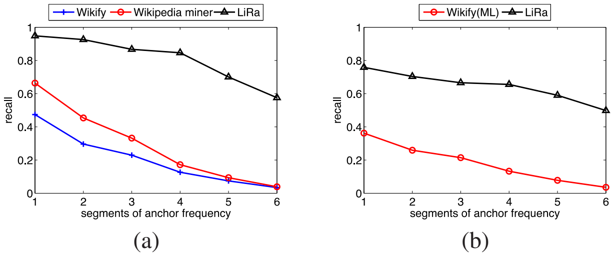
	1	2	3	4	5	6
Freq. range	>100	51–100	11–50	6–10	2–5	1
# of anchors	116	108	527	482	1,399	2,149
Avg. freq.	271.1	70.1	20.7	6.5	2.6	1

We divide the radiology anchors into different segments based on their frequencies, as listed in Table 9. We evaluate the performance of the three systems on anchor detection and overall link generation for the different segments. The performance of each system on segment *seg* is measured using the *recall* function.

$$recall(seg) = \frac{n_{seg}}{|seg|} \tag{9}$$

where n_{seg} is the number of anchor texts in seg that are correctly recognized in the case of anchor detection, or whose target pages are correctly identified in the case of target finding.

We show the systems' performance of anchor detection in Figure 3(a) and that of target finding in Figure 3(b). Since we do not have access to the intermediate results of the Wikipedia Miner system as discussed in Section 6, we only show the performance of Wikify! (machine learning version) and LiRa on target finding.



(a) (b)

Figure 3: Systems' performance differentiated by anchor texts' frequency. Anchors are ranked according to their frequency of occurrence in the radiology reports. X-axes show the ranks of anchors; Y-axes show the systems' scores on the r most frequent anchors, see Eq. 9. Figure 3(a) shows the anchor detection rate; Figure 3(b) shows the automatic link generation rate.

Both anchor detection and target finding show a general trend: the more frequent the anchor texts, the better the systems' performance. Further, we see that LiRa shows more robust performance compared to other systems with respect to infrequent anchor texts.

We conclude that anchor frequency has an impact on the performance of link generation systems in both anchor detection and target finding. Our results indicate that the link generation systems considered achieve better performance on frequent anchor texts than on infrequent anchor texts.

8. CONCLUSION

We conclude the paper by summarizing answers to the research questions raised in Section 1 and discuss directions for future work.

With respect to RQ1, we find that existing link generation systems trained on general domain corpora do not provide a satisfactory solution to linking radiology reports. The major problem is that medical phrases typically have a more complex semantic structure than Wikipedia concepts.

With respect to RQ2, we used a sequential labeling based approach with syntactic features to anchor detection in order to exploit the syntactic regularity present among medical phrases. We then used a sub-anchor based approach to target finding, in order to resolve the complexity in the semantic structure of medical phrases. Our proposed approach was shown to be effective as evaluated on our test collection.

Further, we found that automatic link generation systems tend to achieve better performance in recognizing and finding targets on frequent anchor texts than on anchor texts with a low frequency. In order to achieve robust performance, it is therefore important that a system is effective when dealing with low frequency anchor texts.

A number of directions are left to be explored in the future. These include linking (from radiology reports) to other resources, such as the MedLink encyclopedia. Next, the techniques can be used to provide background information for terminology-rich documents in other domains, ranging from, say biology to zoology. A third line of generalization concerns a slightly different problem to what we have been considering so far, viz. linking textual documents to suitable ontologies, using Wikipedia as a pivot.

Acknowledgements

This research was supported by the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework

Programme, CIP ICT-PSP under grant agreement nr 250430, the Fish4Knowledge project and the PROMISE Network of Excellence, funded and co-funded by the 7th Framework Programme of the European Commission, grant agreement nr 257024 and nr 258191, respectively, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

9. REFERENCES

[1] J. Allan. *Automatic hypertext construction*. PhD thesis, Cornell University, 1995.

[2] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *JAMIA*, 17:229–236, 2010.

[3] W. Bluestein. *Hypertext versions of journal articles: Computer aided linking and realistic human evaluation*. PhD thesis, University of Western Ontario, 1999.

[4] L. Breiman. Random forest. *Machine Learning*, 45(1):5–32, 2001.

[5] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *TPDL 2011*, 2011.

[6] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[7] N. Collier, C. Nobata, and J.-i. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *COLING '00*, pages 201–207, Morristown, NJ, USA, 2000.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[9] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP '07*, pages 708–716, 2007.

[10] D. Ellis, J. Furner, and P. Willett. On the creation of hypertext links in full-text documents: Measurement of retrieval effectiveness. *J. Am. Soc. Inform. Sci.*, 47(4):287–300, 1996.

[11] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *LinkKDD-2005*, 2005.

[12] C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. Johnson. A general natrual language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.*, 1(2):161–174, 1994.

[13] S. Geva, J. Kamps, and A. Trotman, editors. *INEX' 08*, 2009.

[14] S. Green. Building newspaper links in newspaper articles using semantic similarity. In *NLDB '97*, pages 178–190, 1997.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Expl. Newsl.*, 11:10–18, 2009.

[16] P. Haug, D. Ranum, and P. Frederick. Computerized extraction of coded findings from free-text radiologic reports. *Radiology*, 174(2):543–548, 1990.

[17] P. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, and S. Huff. A natural language understanding system combining syntactic and semantic techniques. In *Ann. Symp. Comp. Appl. in Medical Care*, pages 247–151, 1994.

[18] W. C. Huang, A. Trotman, and S. Geva. The importance of manual assessment in link discovery. In *SIGIR '09*, pages 698–699, New York, NY, USA, 2009.

[19] W. C. Huang, S. Geva, and A. Trotman. Overview of the INEX 2009 link the wiki track. In *INEX '09*, 2010.

[20] V. Jijkoun, M. A. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *AND '08*, pages 23–30, New York, NY, USA, 2008. ACM.

[21] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. In *ACL 2002 workshop on Natural language processing in the biomedical domain*, volume 3, pages 1–8, 2002.

[22] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01*, pages 282–289, 2001.

[23] K.-J. Lee, Y.-S. Hwang, and H.-C. Rim. Two-phase biomedical ne recognition based on SVMs. In *ACL 2003 workshop on Natural language processing in biomedicine*, volume 13, pages 33–40, Morristown, NJ, USA, 2003.

[24] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86*, pages 24–26, 1986.

[25] G. Lidstone. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, 1920.

[26] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[27] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, S6, 2005.

[28] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning semantic query suggestions. In *ISWC'09*, pages 424–440, 2009.

[29] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, 2007.

[30] G. A. Miller and E. B. Newman. Tests of a statistical explanation of the rank-frequency relation for words in written english. *American Journal of Psychology*, 71:209–218, 1958.

[31] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08*, pages 509–518, New York, NY, USA, 2008.

[32] W. Mossberg. New Windows XP feature can re-edit others' sites. *The Wall Street Journal*, 2001.

[33] N. Okazaki. CRFsuite: a fast implementation of conditional random fields (CRFs), 2007. URL http://www.chokkan.org/software/crfsuite/.

[34] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, 1998.

[35] M. Porter. An algorithm for suffix stripping. *Program*, 14(3): 130–137, 1980.

[36] R. Schenkel, F. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *BTW2007: Datenbanksysteme in Business, Technologie und Web*, 2007.

[37] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *JNLPBA '04*, pages 104–107, 2004.

[38] K. Takeuchi and N. Collier. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137, 2005.

[39] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04*, pages 575–582, 2004.

[40] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan. Recognizing names in biomedical texts: A machine learning approach. *Bioinformatics*, 20:1178–1190, 2004.