NAIST-IS-MT1651129

Master's Thesis

Domain Adaptation for Sentence Classification: A Study on Structured Abstract Generation

Liu Xinran

March 8, 2018

Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Liu Xinran

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Assistant Professor Satoshi Nakamura	(Co-supervisor)
Assistant Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Hiroyoki Shindo	(Co-supervisor)
Assistant Professor Hiroshi Noji	(Co-supervisor)

Domain Adaptation for Sentence Classification: A Study on Structured Abstract Generation*

Liu Xinran

Abstract

Domain adaptation is particularly important in natural language processing tasks, where manually creating labeled data for a specific domain is costly. Thus leveraging labeled data of available domains has become essential. In this study, we focus on the transferability of knowledge in sentence-level classification, and specifically, on classifying the functional categories of sentences in a scientific abstract, such as the *background*, *objective*, *method*, *result*, and *conclusion*. Considering the significant differences in both the semantics and the syntactics of sentences between the source (e.g., biological science) and target domains (e.g., computational linguistics), we present a framework to automatically construct mappings in semantics and syntactics across domains, and embed these transformation blocks into a context-aware convolution neural network-based classifier. Our results show that our best performing model improves the best baseline by 40%, which demonstrates the effectiveness of the proposed framework.

Keywords:

Domain Adaptation, Sentence Classification, Structured Abstract, Context-aware, Convolutional Neural Network

*Master's Thesis, Graduate School of Information Science,
Nara Institute of Science and Technology, NAIST-IS-MT1651129, March 8, 2018.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Deep-learning techniques are highly effective when trained with a large amount of labeled data. Such data, however, are not always available, and manually constructing enough training data is costly. It is nevertheless often possible to leverage labeled data in another domain to transfer knowledge to an unlabeled domain. Domain adaptation plays an important role in various natural language processing(NLP) tasks (e.g., SRL, NER, POS, and chunking). Considerable effort has been spent in proposing domain adaptation algorithms to solve these typical tasks for a domain with/without few labeled data by leveraging plenty of labeled data in another domain. In this study, we are primarily interested in the task of sentence-level classification, and specifically, classifying the functional categories of the sentences in a scientific abstract, such as the *background*, *objective*, *method*, *result*, and *conclusion*.

Domain adaptation for sentence-level classification has been prevalently studied recently. For example, the typical task of across-domain sentimental analysis [5], aims at predicting the polarity in the sentiments of a sentence in target domain by learning the domain-invariant features to connect two domains. It works by relying on the fact that two domains (e.g., book reviews and movie reviews) share enough semantic/syntactic features, such as adjectives, adverbs, and similar writing style. However, domain adaptation in sentence-level functionality categorization is not trivial. The first big challenge arises from the very different vocabularies and

completely different sets of named entities between the source (e.g., biological science) and target domains (e.g., computational linguistics), which result in lexical feature-based systems performing poorly. Another challenge lies in the differences in writing styles. For example, biological science mainly focuses on new discoveries, where the verbs `observe`, `explore` and `discover` appear more often in *objective* sentences. In contrast, computational linguistics emphasizes new proposals, and *objective* sentences are more likely to be described using verbs such as `propose`, `introduce` and `provide`. Therefore, in addition to the semantic correspondence, identifying the correspondence between writing patterns is essential to transferring learned features.

1.2 Framework

To overcome these challenges, we first introduce a domain adaptation mechanism based on a transformation technique over distributed word representations to align the semantics across two domains. Here, we assume that sentences with the same functionality will also be similar in semantics. The transformation works in an unsupervised manner, where the mapping is constructed by automatically selecting anchor terms that are invariant in semantics across domains. To build the correspondence in syntactics, we propose using several linguistic features, such as the tense of the sentence, which is more likely to remain stable over functional categories in both domains. In addition, noun and verb phrases are utilized to capture more accurate semantics than those in single-word representations. In addition to local information, we utilize contextual information (e.g., features from the previous and next sentences) of the target sentence. For instance, it is more likely that *result* or *conclusion* sentences will appear after a sentence describing the *method*. To jointly learn both the semantic and the syntactic features, we build our model on a convolution neural network in order to capture more syntactic patterns during the process of convolution.

We conduct experiments on the labeled MEDLINE data set from the abstracts of biological science papers and the unlabeled ACL anthology corpus in the field of computational linguistics. A small data set in the ACL anthology containing 463 sentences is manually annotated for evaluation. According to the results, our best performing approach outperforms the state-of-the-art method by 40%, and the fine-tuned model can achieve 60% accuracy over five categories.

1.3 Main Contributions

To sum up, our contributions are as follows: (1) we propose an effective framework to perform across-domain sentence classification by aligning both semantic and syntactic features; (2) we evaluate the proposed approaches on the unstructured text of MEDLINE and ACL anthology data, which proves the effectiveness of our approach; (3) an important characteristic of our approach is that it is unsupervised *. The proposed methods are also generic enough to be applied for any raw-text datasets.

1.4 Organization

This thesis is organized as follows: Chapter 2 introduce the related work about domain adaptation, sentence-level classification, context-aware CNN and structured abstract generation. In Chapter 3, we will provide the problem statement and explain the details of the methodology of domain adaptation in Chapter 4. Chapter 5 describes the design of classifier. Chapter 6 gives the details of our experiment, like experiment settings, baseline, and the results. Chapter 7 summarizes the work and points out some directions of future work.

*the module fine-tuning is in semi-supervised manner

2 Related Work

2.1 Domain Adaptation

The main challenge in domain adaptation is that data in the source and target domains are distributed differently. Since the training data is drawn from a source distribution and the testing data is drawn from a different target distribution, we cannot expect a large sample of our source data to allow us to build a good target model [2]. In fact, Blitzer et al. showed that realistic differences in domains can cause discriminative linear classifiers to more than double in error.

Several techniques have been investigated to eliminate such differences. Daumé III et al. [9] employed generalizable features across domains to reduce domain divergence with feature augmentation-based learning methods. Some studies focused on unsupervised approach due to the lack of labeled data. Gong et al. [11] proposed a kernel-based domain adaptation method that exploits intrinsic low-dimensional structures in the datasets. In their geodesic flow kernel(GFK) method, a metric was introduced to predict the best source domain which is suited for adaptation to a target domain, without using labeled target data. Moreover, Baktashmotlagh et al. [1] overcomes this challenge by extracting information that is invariant across domains, applying a projection of the data onto a low-dimensional latent space, where the distance between the empirical distributions of the source and target examples is minimized. Similarly to the idea of mapping features from the source to the target domain, our approach aligns semantic and syntactic features using a transformation technique embedded in

a CNN framework. As a practical concept, domain adaptation has been applied in plenty of tasks. Xiao and Guo proposed to adapt sequence labeling systems in NLP from a source domain to a different but related target domain by inducing distributed representations using a log-bilinear language adaptation (LBLA) model [30]. They use the learned representation embedding functions to map the original data into the induced representation space as augmenting features, which are incorporated into supervised sequence labeling systems to enable cross domain adaptability. Their approach was shown to outperform lots of related methods for several traditional NLP tasks like cross-domain POS tagging systems, cross-domain syntactic chunking and named entity recognition systems. Despite of the various of methods for domain adaptation have been proposed, none of them has been utilized on the task of sentence-level classification. To the best of our knowledge, we are the first to tackle the problem of domain adaptation in terms of sentence-level functionality categorization.

2.2 Sentence-Level Classification

Sentence-Level classification has become an important area in NLP. Pang et al. [23] applied three machine learning methods (naive Bayes, maximum entropy classification, and support vector machines) to sentiment classification. However, the results do not perform as well as on traditional topic-based categorization, despite the several different types of features they tried.

Recently, Convolutional Neural Networks (CNN) for sentence classification has been developed into a very important application in NLP. Although CNN was used as a model for image recognition initially, the thought of utilizing layers with convolving filters provided a reference to feature extraction in sentence level tasks. Based on CNN, Shen et al. [27] presented a series of latent semantic models to learn low-dimensional semantic vectors for searching queries and Web documents, and achieved remarkable results

in semantic parsing. Dynamic Convolutional Neural Network (DCNN) [14] was also shown to be a very effective model for the semantic modeling of sentences.

All the above research proved the effectiveness and availability of CNN in sentence representation or modeling. Kim [15] suggested that a pre-trained CNN model can recognize some universal feature, which can be utilized for various classification tasks. Furthermore, learning task-specific vectors through fine-tuning results can achieve better results in specific domain. Their outcomes enlighten us on using CNN as a basic model for cross-domain sentence classification, and finally forming an improved model with transformed embedding layer for cross-domain vocabularies mapping.

2.3 Context-aware CNN

While taking the writing pattern correspondence into consideration, the CNN should be able to detect features of contextual information. The context-aware concept was first introduced by Schilit in ubiquitous computing [26]. The concept now plays an important role in improving the usability of search engines by exploiting all kinds of information, such as user feedback [18], anchor text [16], user profiles [7], and so on. Instead of mining patterns of individual queries which may be sparse, H.Cao et al. [4] summarized queries into concepts. A concept is a group of similar queries.

Recent advances in context-aware CNNs have resulted in significant progress in many computer vision tasks. Vu et al. [29] proposed a pairwise model for an object detection task, training the model using an explicit structured-output surrogate loss with an external inference routine that enables the parameters of the model to be fine-tuned simultaneously. Inspired by their results, we introduce a context-aware CNN into a sentence classification task in NLP, where we regard the writing pattern as a kind of contextual information. In other words, we suggest that the text surrounding the target sentence in different functional categories have patterns that can be

captured by a context-aware CNN. In other words, we suggest that the sequence of sentences from different functional categories have some regular characteristics which can be captured by a Context-aware CNN.

2.4 Structured Abstract Generation

Previous investigations [12] have demonstrated that many practical tasks require accessing specific types of information in the scientific literature, such as information about the objective, methods or results. The information structure, that is, the way speakers construct sentences to present new information in the context of old information, can capture rich linguistic information about the discourse structure of scientific documents for other NLP tasks [8]. Being able to automatically construct a structured abstract will benefit data searches and comparisons in the scientific literature.

3 Problem Statement

We define the task as a domain adaptation problem from a source domain $\mathcal{S}$ to a target domain $\mathcal{T}$. In the source domain, we split the data into l^S labeled sentences $\{(X_i^S, Y_i^S)\}_{i=1}^{l^S}$, where X_i^S is the ith input sentence, that is, a sequence of words/phrases, w_1,w_2...,w_j, and Y_i^S is the corresponding label of the sentence, and u^S unlabeled sentences $\{(X_i^S)\}_{i=l^S+1}^{n^S}$, where $n^S = l^S + u^S$. Owing to the lack of labeled data in the target domain, a very small number of labeled sentences are available, $\{(X_i^T, Y_i^T)\}_{i=1}^{l^T}$ ($l^T << l^S$), with the remainder u^T being unlabeled sentences $\{(X_i^T)\}_{i=l^T+1}^{n^T}$, where $n^T = l^T + u^T$.

Because our approach works in an embedded space, we define the source embedding space as E^S with the vocabulary set V^S. The corresponding target embedding space is defined as E^T with the vocabulary set V^T. In our approach, in order to adapt the differences across two domains, we construct a virtual embedding space E^{S+T}, which contains the embedded vocabularies from the source space and the transformed embeddings of vocabularies from the target space (the details of the transformation are provided in Section 4.1.

4 Domain Adaptation

4.1 Semantic Correspondence

Transformation based on Anchor Mapping

In order to represent term semantics, we apply word embedding techniques. A distributed representation of words using neural networks was originally proposed by Rumelhart in 1985 [25], and was later improved upon by Mikolov et al. [19, 20], who introduced the Skip-gram and CBOW models relying on a simplified neural network architecture to generate vector representations of words from text. In order to decrease the semantic gap across two domains, we apply the transformation technique proposed by Zhang et al. [31,32] to construct the mappings of terms in two domains. As mentioned in their work [31,32], the terms in two different semantic vector spaces cannot be compared directly because the features/dimensions in the two spaces have no direct correspondence as a result of their separate training processes. Therefore, a transformation matrix is trained to establish the connection between the vector spaces.

To better understand the concept behind the transformation, one could compare the semantic spaces to buildings. If two semantic spaces are imagined as two buildings, to map the components of one building to the ones in the other one, we need first to learn how the main frames of the two buildings correspond to each other. After this is known, the remaining components can be automatically mapped by considering their relative positions to the main frames of their buildings. The concept behind the transforma-

tion is to map the terms in one semantic space to those in another space. Thus, we first need to learn how the "main frames" of the two semantic spaces correspond to each other. Here the "main frames" are known as the anchor terms across two domains. Once the correspondence between the anchor terms in the two semantic spaces is established, we can automatically map any terms relative to these anchor terms. In this work, we use *shared frequent terms* as anchors to construct the transformation. However, manually preparing sufficiently large sets of anchor terms that cover various aspects and that exist in any possible combinations of the source and target domains requires much effort and resources. Here, we rely on an approximation procedure that automatically proposes anchor pairs. Specifically, we select terms that have high frequency (e.g., `method`, `experiment`, `propose`) in both the source and target spaces. Intuitively, terms that occur frequently in both spaces are more likely to have a stable meaning and are likely to co-occur with many other terms. In our implementation, the top 10%* common frequent terms in the two corpora. This observation has been validated by linguistic studies of several Indo-European languages including English which discovered lower semantic drift of frequently used terms [17, 21]. Even if certain anchor term pairs do not retain the same semantics across domains, the results should not deteriorate significantly when using a sufficiently high number of anchor term pairs.

Suppose there are u pairs of anchor terms $\{(w_1^S,\ w_1^T),\ldots,(w_u^S,\ w_u^T)\}$, where w_i^S is an anchor in one space, and w_i^T is its counterpart, that is, the same anchor in the other space. The transformation matrix $\mathbf{M}$ is then constructed by minimizing the differences between $\mathbf{M}\mathbf{w_i^S}$ and $\mathbf{w_i^T}$ (see Eq. 4.1). In particular, we minimize the sum of the Euclidean 2-norms between the anchor terms in the source domain and their counterpart anchors in the target domain. Eq. 4.1 is applied to solve the regularized least squares problem ($\gamma = .02$), together with a regularization component to prevent over-fitting:

*This number has been experimentally found to perform best.

$$\mathbf{M} = \underset{\mathbf{M}}{\operatorname{argmin}} \sum_{i=1}^{u} \left\| \mathbf{M}\mathbf{w}_\mathbf{i}^\mathbf{S} - \mathbf{w}_\mathbf{i}^\mathbf{T} \right\|_2^2 + \gamma \left\| \mathbf{M} \right\|_2^2, \tag{4.1}$$

where u denotes the size of the anchor term set that contains, in our experiment, we selected the top 10% frequent terms in the intersection of the vocabularies of the two corpora.

4.2 Syntactic Correspondence

Linguistic features have been considered as essential measures to calculate the similarity of sentences. In our study, we assume that sentences that belong to same functional category will also be similar in their linguistic patterns.

Noun and Verb Phrases.

Usually, the concepts in scientific papers are composed of several words. However, the meanings of the phrases differ greatly from those words when they are separated. For instance, "machine translation," which is a frequently mentioned phrase in the NLP domain, has very a different meaning from "machine" or "translation." Similarly, a verb phrase occupies a key place in a sentence. Some verb phrases can imply the function of the sentence explicitly. For example, the verb phrase "figure out" can indicate that this sentence is more likely to be used to describe the objective of the study. Recognizing these phrases accurately and treating them as a unit during embedding can contribute to more precise semantics, and is more likely to identify the connection between these concepts and the functionality of the sentences they form[†].

[†]In the experiments, high frequency expressions are recognized by the bi-gram model, to increase the coverage of noun and verb phrases.

Tense of the Sentence.

Tense as a distinguishing feature, has a strong indication of the functionality of a sentence. In the scientific literature, authors are more likely to use the simple present tense to describe the objective of their study, while applying past tense and perfective tense when illustrating the method or explaining the experiments. These features can be captured not only in abstract section, but also in other parts of scientific papers. Furthermore, these characteristics can be shared across domains, because grammatical norms and the logic of writing design are relatively uniform in normalized articles. In our research, the tense of a sentence is defined as the tense of the dominant verb in the sentence. The verb found in the root of the parse tree is recognized as the dominant verb. We form the parse tree utilizing the Stanford Parser [6, 28].

4.3 Sequential Correspondence

The sequence of functional sentences in an abstract depicts the schematic organization of the scientific paper, to a certain extent, which can be found universally across different domains in scientific papers. These papers usually start with *background* sentences by pointing out the issues on which the paper focuses. This is followed by the *objective*, and then an explanation of the proposed solution, such as algorithms or models. These explanatory sentences can be labeled as the *method*. Next, the *result* is presented to describe the findings based on proposed algorithms. The *result* may also follow the introduction, in the form of a preview, or appear with the *conclusion* of the paper in the form of a summary. Finally, most papers' abstract close with CONCLUSIONS sentences. Although CRFs are commonly used to incorporate the sequential information if regarding an abstract as a sequence of sentences [13], they may learn too much about the sequential ordering of the sentences since the training data (i.e., structured abstracts)

have very consistent ordering in our case which might not be guaranteed in target domain. Therefore, in our study, we propose a context-aware CNN to learn sequential features by considering the features of adjacent sentences (see Section 5.3).

5 Classifier

Based on the features described in the previous section, we propose two classification models, based on a CNN structure, but embedded with different feature sets. The basic version contains (1) a transformation block, constructed using the semantic correspondence across the source and target embedding spaces by mapping them onto a virtual embedding space, which is later added to the embedding layer in the CNN framework; and (2) the syntactic correspondence, constructed by pre-training a phrase embedding for each space, and adding an embedding layer for the tense of the sentence. The context-aware-CNN model extends the basic framework by building three CNN blocks to capture the contextual information from adjacent sentences (previous, target, and next sentences) and then passes the concatenated information to the classifier.

5.1 Transformation Block

The proposed model, shown in Figure 5.1, is based on the CNN architecture introduced by Kim [15], particularly for the task of sentence classification. To adapt this structure to our problem, we introduce a transformation block within the embedding layer, which helps to map the vocabularies from two domains to a virtual embedding space where words can be compared directly. That is, the closer the two terms are in the virtual space, the more similar they are in semantics. In the embedding layer, we treat the transformation block as an off-line process, where we pre-train the word/phrase embedding spaces for each domain. Then, we leverage the transformation

technique introduced in Sec.4.1 to learn a transformation matrix. Thus, the embedding layer (virtual embedding space (E^{S+T})) contains the embedding of the vocabularies from the source space and the transformed word/phrase embeddings from the target space.

5.2 Convolutional Neural Network (CNN) Model

The input of the model is a sentence of length n words (padded, where necessary), where each word is represented by a k-dimensional vector in the virtual embedding space, and then concatenated as the input to the convolutional layer. Multiple filters with different widths are applied to a window of words, with a convolution operation to generate different feature maps (e.g., a filter with width of three convolves a window of any three continuous words in the input sentence into a new feature). Then, a max-pooling operation is applied to obtain the maximum value, considered the most important feature, in each feature map.

Because the tense information is extracted for each sentence, rather than for each word in the sentence, we treat the tense as an auxiliary feature by concatenating it with the value obtained from the max-pooling. Then, we pass the concatenated feature to a fully connected softmax layer that outputs the probability distribution over the labels. To enable the distributed representation of tense, a random initialized embedding layer is added to the auxiliary input.

Both the pre-trained vectors in the semantic embedding layer and the random initialized tense embedding layer are fine-tuned using back propagation to cater to the final objective function in the classification.

Figure 5.2 shows the procedures of training and testing. In the training procedures, we firstly construct a phrase embedding space for Bio-Science by word2vec, update the embedding after training by CNN classifier. In the

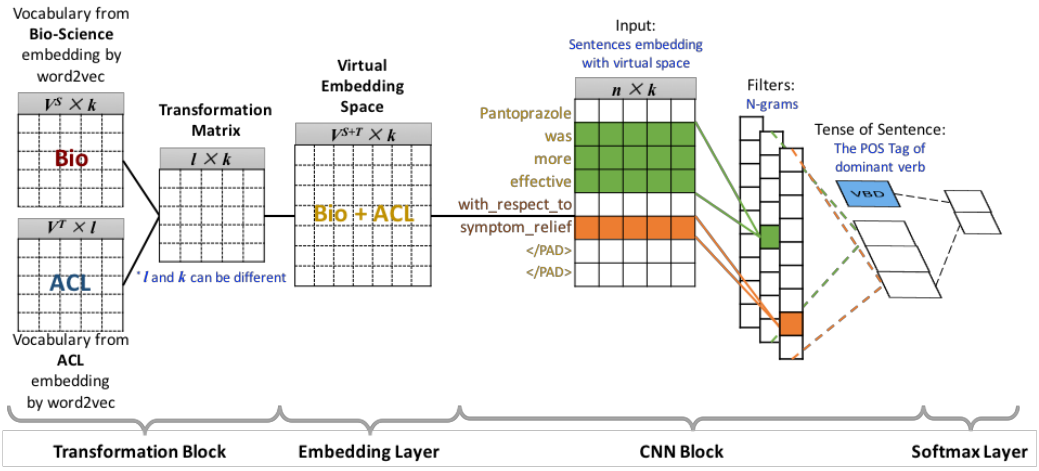

Figure 5.1: Outline of the classifier in the CNN model.

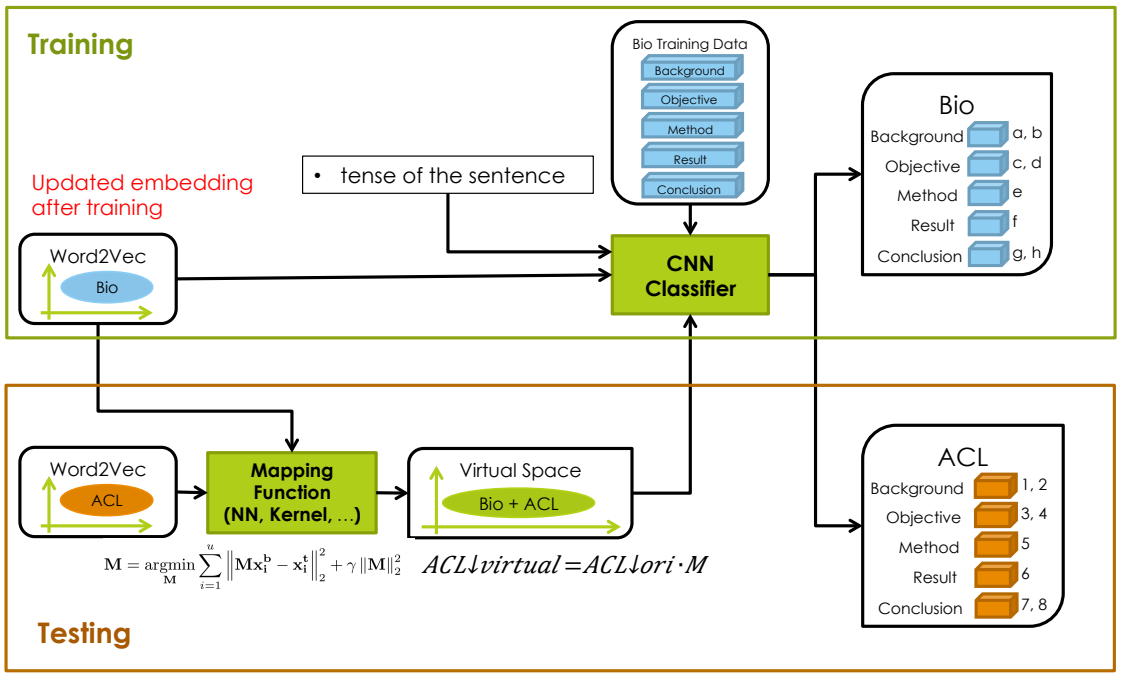

Figure 5.2: The training and testing process.

testing procedures, we calculate a transformation block through mapping function, map the terms in target domain onto a virtual embedding space, which is later added to the embedding layer in the CNN framework;

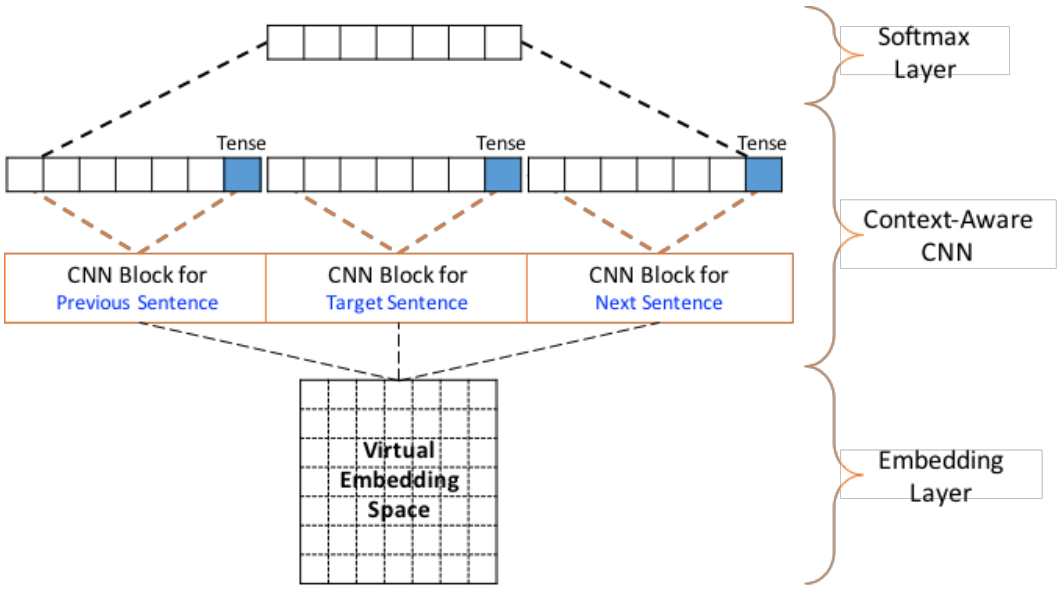

Figure 5.3: Outline of the classifier in the context-aware CNN model.

5.3 Context-aware CNN Model

A context-aware CNN is an extended version of the CNN model introduced above. The main idea of this model is to take into consideration the contextual information of the target sentence in order to adapt the assumed correspondence of writing patterns across two domains.

The architecture of the context-aware CNN model is depicted in Figure 5.3, where another two CNN blocks are added for the previous and the next sentence of the target sentence. The three CNN blocks share one embedding layer, as in the CNN model. Each CNN block contains a convolutional layer, a pooling layer, and a tense layer, where all the settings are the same in terms of the filters (number and width), max-pooling operations, and tense embedding space. Then, the three sets of features from each CNN block are merged by a concatenation operation and passed to the softmax layer for prediction. The parameters in the embedding layers are updated in the same manner as in CNN model.

6 Experiments

6.1 Experimental Settings

In this section, we describe our datasets and explain the setting of our experiment. In the following, we show our results in later part of this section and give some discussions.

6.1.1 Data Set

We perform an extensive evaluation of the proposed approach on two collections of scientific literature from the domains of biological science (MEDLINE corpus) and computational linguistics (ACL anthology corpus [24]). We collected papers from MEDLINE 2014&2015, where structured abstract is employed by labeling the sentence with its role of using one of the following classes: background, objective, method, result, and conclusion. We prepared two training sets, randomly selected from the entire corpus: MEDLINE-small tests the effectiveness of each component in the proposed model; and MEDLINE-large is utilized to compare the proposed model with baseline methods. In the target domain, considering there is no standard benchmark, we selected 100 abstracts containing 463 sentences from the ACL Anthology corpus as a test set, and manually annotated each sentence with the same structural labels as those in the MEDLINE data set. The size and distribution of these annotated data are shown in Table 6.1. In addition, the entire raw text, including both the abstract and the paper content in MEDLINE 2014&2015 (around 10 million sentences

with 336k vocabularies) and in the ACL Anthology (containing 3.7 million sentences with 448k vocabularies) are used to construct pre-trained word embeddings with word2vec [19, 20] for the two domains.

Table 6.1: Statistics of annotated data sets.

		BKG	OBJ	METH	RES	CONC
MEDLINE-small	30,000	3,443	2,515	8,104	11,170	4,768
MEDLINE-large	222,790	20,906	19,138	59,249	86,059	37,438
ACL Anthology	463	88	93	110	147	25

6.2 Baselines

We set up four baselines. **Naive Bayes** is a linear classifier that is known for being simple, yet very efficient. In our multi-class labeling task, we compute the probability of each class label, then pick the class with the largest probability. Bags-of-words in the sentence are used as the feature set. Support vector machines are another supervised learning method often used in classification tasks. Two variants are tested by applying different feature sets: **SVM(BOW)** is set to use bag-of-words features in the sentence, and **SVM(W2V)** utilizes sentence embedding features obtained by averaging the word embeddings of all the words in the target sentence. Domain-adversarial training of neural networks (**DANN**) [10] a simple and effective method for accomplishing domain adaptation with a gradient reversal layer, which ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in choosing domain-invariant features.

6.2.1 Proposed Methods

To verify the effectiveness of four proposed features (the transformation based on anchor mapping (+**Tran**), noun and verb phrases detection using the phrase-gram embedding model (+**bi**), tense of sentence annotation (+**tense**), and sequential correspondence using the context-aware CNN (+**context**)), we conducted experiments over all 16 combinations of feature sets (see Table 6.2). Note that **noTran** stands for the implementation without the transformation block, but using a joint vector space trained on the combined corpus of the two domains. Then, +**uni** represents the method using the unigram embedding model. The best performing combinations with +**noTran** and +**Tran** are later compared with against the baseline methods (see Table 6.3 and 6.4).

6.3 Experimental Results

Table 6.2 shows the evaluation of the accuracy of the proposed model with different combinations of feature sets. The main observation is that the four proposed feature components proved to be beneficial for transferring learning in the sentence classification task. Table 6.3 and 6.4 displays the comparison between our best performing methods and the baselines with respect to the performance multiple tasks.

Table 6.3 shows the results of (1) the supervised sentence classification in a single source domain (columns "Sup.(S$\longrightarrow$S)") and target domain ("Sup.(T$\longrightarrow$T)") where a 5-fold cross validation is conducted for each test; (2) the unsupervised domain adaptation in the sentence classification (column "Unsup.(S$\longrightarrow$T)") by training with only the source domain data and validating on the target domain data; (3) the semi-supervised domain adaptation (column "Semi-Sup.(S+T$\longrightarrow$T)") where the four baselines are trained with combined data from the source domain and part of data from the target domain, the "S+T" represents the combined dataset.

In our proposed method, we first pre-trained the model on source domain data and fine-tuned with the part of target domain data. In current implementation, 80% of labeled sentences (370 sentences) in ACL anthology test set are used during semi-supervised learning process and the rest 20% (93 sentences) are left for evaluation. Note that the "S+T" here represents a fine-tuned results instead of combined dataset. The results are obtained through a 5-fold cross validation on the test set. We discuss the results in detail below.

More specifically, Table 6.4 shows the accuracy of different category in unsupervised domain adaptation in the sentence classification.

6.3.1 Necessity of Transformation

As mentioned before, the tested approach **+noTran** does not contain a transformation block on the embedding layer. Instead, a joint embedding space trained on the combined corpus of MEDLINE and the ACL Anthology is utilized for word/phrase representations. Compared to the methods without transformation (**+noTran**), the methods with transformation block (**+Tran**) improve the overall accuracy by 4%, on average. This proves the effectiveness of using a transformation block to construct semantic correspondence across domains. More specifically, it enforces the words/phrases from two domains with similar functionalities, but with different literal forms, to be closer to each other in the virtual embedding space. As a result, the features for classification learned from the source domain can be better adapted/applied to the target domain. For the category *background*, the better performance of **+noTran** might indicate that there exist more general, or the same terms in the *Background* of the two domains, which would mean that the joint learning of word/phrase embeddings can better capture the domain-invariant features.

Table 6.2: The accuracy of the results over different combinations of the proposed feature sets. The MEDLINE-small data set is used as the source domain.

	Methods	**Overall**	**BKG**	**OBJ**	**METH**	**RES**	**CONC**
1	noTran-uni	0.253	0.398	0.269	0.200	0.122	0.680
2	noTran-bi	0.285	0.557	0.237	0.255	0.143	0.480
3	noTran-uni-tense	0.274	0.364	0.323	0.236	0.143	0.720
4	noTran-bi-tense	0.307	0.511	0.355	0.245	0.143	0.640
5	noTran-uni-context	0.309	0.591	0.333	0.191	0.109	0.920
6	noTran-bi-context	0.303	0.682	0.355	0.141	0.075	0.720
7	noTran-uni-tense-context	0.302	0.602	0.355	0.200	0.075	0.840
8	noTran-bi-tense-context	0.330	0.693	0.387	0.227	0.082	0.760
	Average (noTran.)	*0.295*	*0.550*	*0.327*	*0.212*	*0.111*	*0.720*
9	Tran-uni	0.261	0.432	0.366	0.118	0.122	0.720
10	Tran-bi	0.302	0.398	0.183	0.282	0.238	0.880
11	Tran-uni-tense	0.287	0.398	0.194	0.273	0.238	0.600
12	Tran-bi-tense	0.315	0.466	0.430	0.227	0.143	0.760
13	Tran-uni-context	0.302	0.443	0.387	0.209	0.156	0.760
14	Tran-bi-context	0.320	0.500	0.301	0.355	0.143	0.640
15	Tran-uni-tense-context	0.309	0.455	0.419	0.236	0.129	0.760
16	Tran-bi-tense-context	**0.354**	0.614	0.441	0.355	0.088	0.680
	Average (Tran.)	*0.306*	*0.463*	*0.340*	*0.257*	*0.157*	*0.725*

Table 6.3: The accuracy of the baselines and the proposed best-performing methods. The MEDLINE-large data set is used as the source domain. "Sup." and "Unsup." indicated supervised and unsupervised classification task. The left side of the arrow represents the training dataset and the right side represents the test dataset, "S" and "T" are abbreviation for "source domain" and "target domain". For instance, the columns "Sup.(S⟶S)" shows the results of supervised classification, which is trained with dataset from the source domain and test on dataset form same source domain

Methods	**Sup. (S⟶S)**	**Sup. (T⟶T)**	**Semi-Sup. (S+T⟶T)**	**Unsup. (S⟶T)**
naive Bayes	0.762	0.337	0.427	0.246
SVM (BOW)	0.772	0.500	0.466	0.289
SVM (W2V)	0.573	**0.533**	0.379	0.283
DANN	-	-	0.318	0.279
noTran-best	**0.820**	0.453	0.541	0.376
Tran-best	-	-	**0.592**	**0.406**

Table 6.4: The accuracy of different category in unsupervised domain adaptation in the sentence classification.

Methods	**Overall**	**BKG**	**OBJ**	**METH**	**RES**	**CONC**
naive Bayes	0.246	0.250	0.376	0.236	0.050	0.920
SVM (BOW)	0.289	0.295	0.312	0.318	0.177	0.720
SVM (W2V)	0.283	0.068	0.150	0.409	0.374	0.440
DANN	0.279	0.250	0.000	0.000	0.728	0.000
noTran-best	0.376	0.716	0.441	0.282	0.136	0.760
Tran-best	**0.406**	0.682	0.462	0.372	0.163	0.800

6.3.2 Phrase-gram has Better Representation Power in the Embedding Space

As mentioned in Section 4.2, the semantics of a phrase might be quite different to that of a single word within the phrase (e.g., figure out vs. figure, machine translation vs. machine). Therefore, theoretically, phrase-gram embeddings would capture more precise semantics than uni-gram embeddings would. The results shown in Table 6.2 show that the methods with **+bi** outperform those with **+uni** by 10%, on average, which verifies the high representation power of phrase-gram in the embedding space.

6.3.3 The Tense Feature is Effective in Domain Adaptation

Comparing the results of the methods containing the tense feature with the corresponding methods without the feature, the former help to improve the classification performance by 6.3%, on average. To a certain extent, this proves the validity of the tense feature. However, we observed that as the dataset grow, the effectiveness of tense gradually decreased, which may due to the diversity of writing style in various sub-domain.

6.3.4 Advantage of using Contextual Information

Owing to the advantage of the context-aware model that captures the sequential information among sentences, the classification accuracy increases by 11%, on average. By analyzing their confusion matrices across different categories, the contextual information helps to avoid mis-classifying the sentences in the categories of *background*, *objective*, and *conclusion* because of their similar semantics and syntactics.

6.3.5 Improvement over the State-of-the-Art Methods

Table 6.3 and 6.4 compares the performance of our best-performing methods and the baselines over different tasks. We show the performance of the proposed classifier in the typical task of sentence classification*. The proposed model **noTran-bi-tense-context** can achieve 0.82 in accuracy, which outperforms the best-performing baseline **naive Bayes** (see column "Source"). In the unsupervised domain adaptation task, which is discussed in some detail in this paper (see column "Overall"), our model **Tran-best** (or **Tran-bi-tense-context**) outperforms the best baseline by 40%. We also notice the poor performance of the baseline **DANN**, which indicates that the domain-invariant features learned by adversarial training are not enough for the task of across-domain functionality categorization and those domain-specific but semantic corresponding features are indispensable. The semi-supervised learning results are shown in column "Joint", where the proposed model **Tran-best** achieves 0.592 in terms of accuracy. When only using 80% of the labeled ACL sentences for training, without domain adaptation, the highest accuracy 0.533 is achieved by the baseline **SVM(W2V)** (see column "Target"), which demonstrates the necessity of domain adaptation when limited training data are available in the target domain.

*The data sets for training and testing come from the same domain, and there exist enough labeled data for training.

7 Future work: Abstract Generation from main text

7.1 Abstract Generation Model

Since the effectiveness of the methods has been demonstrated, we apply our model to generating a structured abstract by extracting the most representative sentence in each functional category from the main text of the paper. We choose the fine-tuned model that trained by jointed data for its better performance.

The outline of abstract generation model is illustrated in Figure 7.1. Firstly, we acquire the probability distribution of classification for each sentence in main text of a paper from our proposed CNN model. Then we calculate the skewness for each sentence in each category, and give 5 ranking results. By selecting the top sentence in the categories of *background*, *objective*, *method*, *result*, and *conclusion*, we construct a structured abstract, which contains the primary functional information of this paper.

7.2 Evaluations

The evaluation of automatic abstract generation becomes an inevitable problem. Typically, there are two way to evaluate the generated abstract, which are intrinsic and extrinsic methods. In intrinsic method, a reference summary is provided to evaluate the quality of the generated abstract. The

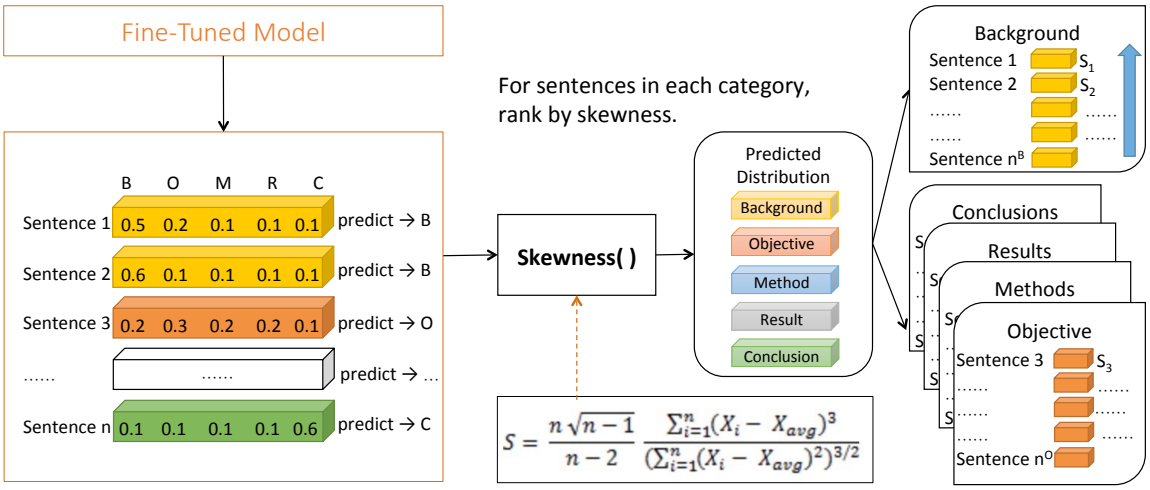

Figure 7.1: Outline of Abstract Generation model.

more identical between generated abstract and the reference summaries, the higher the quality of the generated abstract is. In extrinsic methods, we use the generated abstract instead of the original paper to perform a document related applications like document retrieval, document clustering, document classification. If the alternative abstract can improve the performance, the abstract generation is considered to achieve good qualities.

For preliminary evaluation, we generated some ground truth as reference summary. We annotated some ground truth in the main text of a paper by estimating the similarity between sentences in original abstract and main text, and selecting the most similar ones, considering these are the reference summary.

We implement three method to calculate the similarity between two sentences: TF-IDF, Sent2Vec. Table 7.1 shows some examples of preliminary result. On some cases, the two measures can reach agreement on the most similar sentence form contents. In this example, for the method sentence in abstract, both TF-IDF and Sent2Vec find a content sentence from the method section, which seems have a similar meaning. For the other sentences on the Top-3 sentence, TF-IDF found some “results“ sentences which mainly contain similar words with the abstract sentence while Sent2Vec found sentences all from “discussion“ section which seems less reasonable, for just containing a few keywords.

Table 7.1: Evaluation

Source	Label or Section	Sentence
Original abstract	METH	Coverage data obtained were validated and considered for inclusion in the PCT databank in a collaboration between UNICEF and WHO.
TF-IDF	1. Method	The coverage estimates collected by UNICEF were considered for updating the initial 2013 global coverage data in the PCT databank.
	2. Results	Of the 73 data points in 39 countries with validated data in the UNICEF deworming database, 24 data points (33%) were excluded from analysis because the number of pre-SAC reached and targeted was considered to exceed plausible estimates of demographic data using WHO criteria.
	3. Results	Of the 68 points in the initial WHO PCT databank, 14 (21%) data points were identical to points in the UNICEF deworming database and were chosen in favor of their UNICEF deworming database counterparts, 49 (72%) points were included in the updated PCT databank while 5 (7%) points were excluded from analysis as they were replaced with data points from the UNICEF deworming database.
Sent2Vec	1. Method	The coverage estimates collected by UNICEF were considered for updating the initial 2013 global coverage data in the PCT databank.
	2. Discussion	Prior to this coverage assessment, global pre-SAC deworming was inadequately covered in the official 2013 PCT database.
	3. Discussion	The design of CHDs and the package of interventions offered can be tailored to the local contexts.

8 Conclusion

In this study, we explore the transferability of knowledge in the task of sentence-level classification. Specifically, we use the concept of an information structure model to classify the functional categories of the sentences in the abstracts of scientific papers. Considering the significant gap in both the semantics and the syntactics of the sentences from the source domain and the target domain, we present a framework to automatically construct the correspondence in semantics and syntactics across domains, and embed the transformation block into a context-aware CNN-based classifier. The results show that our proposed framework outperforms the baselines and achieves better accuracy in comprehensive experiments, demonstrating the effectiveness of our methods. We would like to explore more applications of this method in the future, which may contains abstract generation form main text of scientific papers or paper recommendation. From another perspective, these work can be an inspiration of detecting regular patterns of structural organization according to specific topic, or knowledge discovery through structural document understanding.

Acknowledgements

I would like to express my gratitude to all those who helped me during the writing of this thesis. I gratefully acknowledge the help of my academic advisor and supervisor, Professor Yuji Matsumoto, who has offered me valuable suggestions in the academic studies. During my master study, he showed excellent guidance, extraordinary patience and caring. In the preparation of the thesis, he has spent much time reading through my draft and provided me with inspiring advice. Without his patient instruction, insightful criticism and expert guidance, the completion of this thesis would not have been possible.

I also owe a special debt of gratitude to Dr.Zhang Yating from RIKEN AIP Center, from whose devoted teaching and enlightening ideas I have benefited a lot and academically prepared for the thesis.

I would like to thank all my other tutors and teachers in NAIST, for their direct and indirect help to me, and thanks to the Japanese Ministry of Education, Culture, Sports, Science, and Technology for providing me support through MEXT Scholarships.

I should finally like to express my gratitude to my beloved parents who have always been helping me out of difficulties and supporting and encouragement without a word of complaint. I also owe my sincere gratitude to my friends and my fellow classmates who gave me their help and time in listening to me and helping me work out my problems during the difficult time.

References

[1] BAKTASHMOTLAGH, Mahsa, et al. Unsupervised domain adaptation by domain invariant projection. In: Proceedings of the IEEE International Conference on Computer Vision. 2013. p. 769-776.

[2] BLITZER, John, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: ACL. 2007. p. 440-447.

[3] BORGWARDT, Karsten M., et al. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 2006, 22.14: e49-e57.

[4] CAO, Huanhuan, et al. Context-aware query suggestion by mining click-through and session data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008. p. 875-883.

[5] CHEN, Xilun, et al. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. arXiv preprint arXiv:1606.01614, 2016.

[6] CHEN, Danqi; MANNING, Christopher. A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 740-750.

[7] CHIRITA, Paul-Alexandru; FIRAN, Claudiu S.; NEJDL, Wolfgang. Personalized query expansion for the web. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007. p. 7-14.

[8] CONTRACTOR, Danish; GUO, Yufan; KORHONEN, Anna. Using Argumentative Zones for Extractive Summarization of Scientific Articles. In: coling. 2012. p. 663-678.

[9] DAUMÉ III, Hal. Frustratingly easy domain adaptation. arXiv preprint arXiv:0907.1815, 2009.

[10] GANIN, Yaroslav, et al. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016, 17.59: 1-35.

[11] GONG, Boqing, et al. Geodesic flow kernel for unsupervised domain adaptation. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012. p. 2066-2073.

[12] GUO, Yufan, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2010. p. 99-107.

[13] J. Lafferty, A. McCallum, F.Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML. 2001.

[14] KALCHBRENNER, Nal; GREFENSTETTE, Edward; BLUNSOM, Phil. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.

[15] KIM, Yoon. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[16] KRAFT, Reiner; ZIEN, Jason. Mining anchor text for query refinement. In: Proceedings of the 13th international conference on World Wide Web. ACM, 2004. p. 666-674.

[17] LIEBERMAN, Erez, et al. Quantifying the evolutionary dynamics of language. Nature, 2007, 449.7163: 713.

[18] MAGENNIS, Mark; VAN RIJSBERGEN, Cornelis J. The potential and actual effectiveness of interactive query expansion. In: ACM SIGIR Forum. ACM, 1997. p. 324-332.

[19] MIKOLOV, Tomas, et al. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111-3119.

[20] MIKOLOV, Tomas, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[21] PAGEL, Mark; ATKINSON, Quentin D.; MEADE, Andrew. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature, 2007, 449.7163: 717-720.

[22] PAN, Sinno Jialin, et al. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 2011, 22.2: 199-210.

[23] PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002. p. 79-86.

[24] RADEV, Dragomir R., et al. The ACL anthology network corpus. Language Resources and Evaluation, 2013, 47.4: 919-944.

[25] RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[26] SCHILIT, Bill; ADAMS, Norman; WANT, Roy. Context-aware computing applications. In: Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on. IEEE, 1994. p. 85-90.

[27] SHEN, Yelong, et al. Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014. p. 373-374.

[28] TOUTANOVA, Kristina, et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003. p. 173-180.

[29] VU, Tuan-Hung; OSOKIN, Anton; LAPTEV, Ivan. Context-aware CNNs for person head detection. In: Proceedings of the IEEE International Conference on Computer Vision. 2015. p. 2893-2901.

[30] XIAO, Min; GUO, Yuhong. Domain adaptation for sequence labeling tasks with a probabilistic language adaptation model. In: International Conference on Machine Learning. 2013. p. 293-301.

[31] ZHANG, Yating, et al. Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. In: ACL (1). 2015. p. 645-655.

[32] ZHANG, Yating, et al. The past is not a foreign country: Detecting semantically similar terms across time. IEEE Transactions on Knowledge and Data Engineering, 2016, 28.10: 2793-2807.

Publication List

[1] Yating Zhang, Xinran Liu, Yuji Matsumoto. Domain Adaptation for Sentence Classification: A Study on Structured Abstract Generation. SCIDOCA, 2017.