Kernel Fuzzy C-Means Clustering for Word Sense Disambiguation in BioMedical Texts

K. REN[1,2*], Y.F. REN[3]
[1]Computer School, Wuhan University
Wuhan, 430072, China
[2]College of Computer Science
South-Central University for Nationalities
Wuhan, China
[3]Singapore University of Technology and Design
487372, 8 Somapah Road, Singapore
rk8123@gmail.com

JDIM
Journal of Digital Information Management

ABSTRACT: *Word sense disambiguation (WSD) in biomedical texts is important. The majority of existing research primarily focuses on supervised learning methods and knowledge-based approaches. Implementing these methods requires significant human-annotated corpus, which is not easily obtained. In this paper, we developed an unsupervised system for WSD in biomedical texts. First, we predefine the number of senses for an ambiguous word. Kernel fuzzy C-means clustering is used to group the same sense terms into a set. Each set is mapped to a certain sense to achieve disambiguation. Experimental results on all 50 ambiguous terms from NLM-WSD corpus demonstrate that our proposed system outperforms other unsupervised methods. Meanwhile, the kernel fuzzy C-means system is 5% more precise than the state-of-art knowledge-based WSD system on the full NLM-WSD dataset. Our system is highly efficient and accurate for word sense disambiguation in biomedical texts and does not require human-annotated corpus.*

Subject Categories and Descriptors
I.2.7 [Artificial intelligence]: Natural Language Processing - Text analysis; **I.5.3 [Pattern Recognition]**: Clustering - Algorithms

General Terms: Algorithm, Performance

Keywords: Clustering, Kernel fuzzy C-means, Unsupervised Learning, Word sense disambiguation

Received: 4 August 2015, Revised 10 September 2015, Accepted 20 September 2015

1. Introduction

Word sense disambiguation (WSD) aims to select the right sense of an ambiguous word. WSD is a preceding work for a number of bioinformatics tasks, such as name entity reorganization, protein–protein interaction extraction [1], and biomedical information retrieval. Previous works on WSD in the biomedical domain are generally categorized in three classes: supervised methods, knowledge-based approaches, and unsupervised systems.

Supervised methods are usually based on learning from human-annotated corpus. We can obtain results by the annotated corpus in the machine learning framework. After training, the algorithm will be used to run the test corpus. The main disadvantage of these methods is its reliance on human-annotated data. However, obtaining annotated data is expensive. Biomedical texts are huge data that increase all the time. Thus, expanding this type of method to all WSD terms in biomedical texts is impractical.

Knowledge-based methods rely on external knowledge resources, such as dictionaries, Wikipedia and WordNet. In the biomedical domain, the most common knowledge resource is the Unified Medical Language System (UMLS). UMLS is a large system containing a vast range of information in the biomedical domain. Extracting useful information from UMLS is the key step. Biomedical literature is updated daily. Thus, the UMLS cannot cover all the information that individuals need. For this reason, the final result of knowledge-based methods is affected by the

quality of the external resources. This type of method lacks general applicability.

The last type of method is unsupervised approaches. Unsupervised methods do not require annotated corpus for training. A typical example is graph-based methods. Graph-based methods can apply to the latest unlabeled corpus. Thus, we focus on the unsupervised system for biomedical word sense disambiguation. Our system, which is based on kernel fuzzy C-means clustering, only requires the total number of senses to be determined in advance. In the biomedical domain, we can adopt the concept unique identifier (CUI) numbers from the UMLS as the sense number. An unsupervised kernel fuzzy C-means algorithm is employed to cluster terms in the same sense. Finally, we connect each set to different CUIs to complete the WSD task.

Using kernel functions reduces computational complexity and provides the following advantages: Kernel Fuzzy C-means exhibit higher accuracy compared with other unsupervised systems. In a subset of NLM–WSD for biomedical word sense disambiguation, the KFC method achieves an accuracy of 0.64, which is 3 points higher than the SENSATIONAL [2] unsupervised system. Agirre et al. [3] developed a classical unsupervised system on the full NLM–WSD data set, which based on the Personalized PageRank (PPR) algorithm, and achieved an average accuracy of 0.67. The accuracy of our system at the same dataset is 0.81, which is 14% higher than that of the PPR system. Our KFC method can cluster the context in the level of $O(N^2)$. Thus, this approach can be used in a large data set. The algorithm can run in an ordinary PC and be completed in an acceptable duration.

2. Related Works

Clustering algorithm, which is used for word sense disambiguation, is based on the following linguistic theory: Ambiguous words with the same sense have similar contexts.

Supervised methods are the most popular approaches in the biomedical WSD domain. Ginter et al. [4] utilized supervised SVM machine learning method based on the weighted bag-of- words. Their approach increased in accuracy from 79% to 82%. Liu et al. [5] tested a subset of NLM-WSD corpus by integrating their method with the Naïve Bayes. Leroy and Rindflesch [6] proposed a supervised WSD method that maps the ambiguous words into UMLS CUIs. This method extensively affected biomedical word sense disambiguation. Joshi et al. [7] conducted an experimental comparison on a subset of NLM–WSD corpus based on the Naïve Bayes, SVM, AdaBoost, and Decision Tree methods. Their research demonstrated that the SVM-based supervised methods could yield better results. Savova et al. [8] investigated a WSD system by applying a variation of Huber's algorithm. The researchers ran experiments on 28 feature sets and achieved the average F-score of 0.86 on the full NLM–WSD corpus. Stevenson et al. [9] proposed a supervised method, which combined three methods, and tested this approach on the full NLM–WSD corpus. Hisham Al-Mubaid and al. [10] proposed a method utilizing the mutual information between context words to induce reliable learning models for sense disambiguation. Their approach was competitive.

Although the studies show that supervised approaches can yield better results in the biomedical WSD, these methods require manual labeled data for training, which differs from supervised learning. This requirement limits the application of these approaches. Recent studies adopted knowledge-based and unsupervised approaches to solve these problems. Savova et al. [11] used an unsupervised learning method for biomedical applications. This study confirmed that the similarity of second-order co-occurrence works well in biomedical WSD. Schijvenaars et al. [12] and Pahikkala et al. [13] utilized the large-scale dictionary collected from four databases to assist the WSD task. Liu et al. [14] employed UMLS as ontology to obtain external knowledge. Gaudan et al. [15] used the SVM to execute the abbreviations in WSD. Humphrey et al. [16] adopted Medline as extended training data. The researchers proposed the journal descriptor indexing (JDI) method, which utilized the semantic type of UMLS to disambiguate words. Navigli and Lapata [17], Sinha and Mihalcea [18], and Tsatsaronis et al. [19] used the graph-based algorithm to analyze the meaning of ambiguous terms. Duan et al. adopted Max-margin clustering for word sense disambiguation and achieved significantly enhanced results. Agirre et al. proposed an unsupervised system that disambiguates words by converting the tables from the UMLS into a graph based on the Personalized PageRank algorithm. This method can map best sense into ambiguous words. Jimeno-Yepes et al. [20] compared four approaches. Their method employed the semantic types assigned to the concepts in the Metathesaurus and achieved better results. Jimeno-Yepes et al. [21] developed the MSH–WSD data set that consists of 106 ambiguous abbreviations, 88 ambiguous terms, and 9 combinations of both, producing a total of 203 ambiguous entities. For each ambiguous term or abbreviation, the data set contains no more than 100 instances per sense extracted from MEDLINE.

This article presents an unsupervised kernel-based WSD algorithm that is capable of disambiguating all ambiguous words in the biomedical domain. Our experiment covered the full NLM–WSD corpus and obtained favorable results.

3. Method

Our proposed framework is shown in Figure 1. The framework includes three steps: preprocessing and feature extraction, unsupervised word clustering, and sense mapping. The details of each step are described in the following subsections.

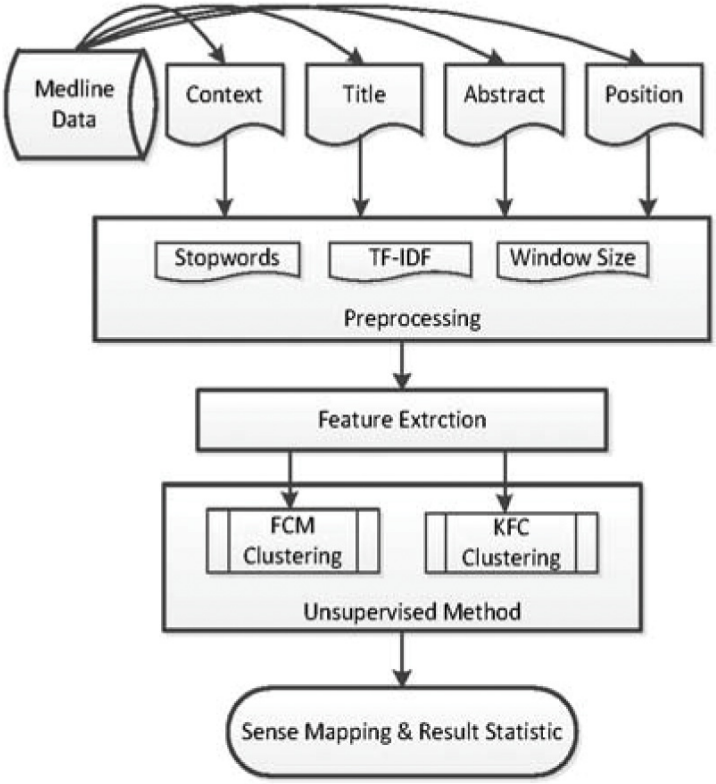

Figure 1. Framework workflow of the proposed WSD system

3.1 Preprocessing and Feature Extracting

First, we filter out all the stop words in the contexts. We compute the frequency of each word and set a threshold to eliminate low-frequency words. We adopt the common tf-idf technique to count the frequency features in the context. The tf-idf feature we used is defined as follows: we let $D = \{d_1, ..., d_n\}$ be the set of documents and $T = \{t_1, ..., t_m\}$ the set of target words occurring in D. We compute the frequency of word $t \in T$ in the document $d \in D$ as $tf(d, t)$. *tf–idf* is a weighting scheme that weighs the frequency of word t in document d with a factor that deducts its importance with its occurrences in the whole document. The definition is presented in Equation (1)

$$tf - idf(d,\ t, D) = tf(d, t) \times idf(t, D) \quad (1)$$

where $df(t)$ is the number of documents where word t appears. Thus, the feature vector representation of document d is defined as Equation (2).

$$\vec{t}_d = tf - idf(d, t_1), ..., tf - idf(d, t_m) \quad (2)$$

We define the size of the context window, that is, the number of words in the context of the words. The appropriate window size that can reserve features is selected. After preprocessing, the contexts of each word are converted into feature vectors.

3.2 Text clustering

In this paper, we use two clustering algorithms for word sense disambiguation: fuzzy C-means (FCM) algorithm and the kernel fuzzy C-means (KFC) algorithm. FCM algorithm is a typical soft clustering method, which is an improvement of the K-means algorithm. FCM algorithm can significantly enhance the performance of K-means clustering. Through the combination of the kernel function, the KFC algorithm maps the linear relationship to the corresponding function space. The features in high-dimensional space can be easily separated. Thus, this algorithm helps improve clustering quality. We will introduce two algorithms in the following subsections.

3.2.1 Fuzzy C-means algorithm

The fuzzy C-means (FCM) was proposed by J. C. Bezdek [22]. In the FCM algorithm, $X = \{X_i\}_{i=1}^{N}$ is a set of N feature vectors. The fuzzy clustering algorithm maps data X into C fuzzy clusters.

$$J(U,V) = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^{m} d^2(X_i, V_c)$$

$$\sum_{c=1}^{C} u_{ic} = 1 \quad \forall i, \ u_{ic} \geq 0 \quad \forall i,c \text{ and } \sum_{i=1}^{N} > 0 \quad \forall c \quad (3)$$

Matrix U represents a fuzzy partition and u_{ic} denotes the degree that X belongs to the cluster C, as shown in Equation (3). In Equation (3), m is the fuzzy degree. Thus, when $m = 1$, the fuzzy C-Means algorithm is equal to the normal k-means algorithm. V_c represents the C cluster centers.

$$J_\lambda(U,V) = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^m d^2(X_i,V_c) + \lambda\left(\sum_{c=1}^{C} u_{ic} - 1\right) \quad (4)$$

The algorithm utilizes the Lagrange multipliers (4) to optimize the parameters.

$$u_{ic} = 1 \Big/ \sum_{c'=1}^{C} \left(\frac{d(X_i,V_c)}{d(X_i,V_{c'})}\right)^{\frac{2}{m-1}} \quad (5)$$

$$V_c = \sum_{i=1}^{N} u_{ic}^m X_i \Big/ \sum_{i=1}^{N} u_{ic}^m \quad (6)$$

The algorithm updates u_{ic} and V_c using (5) and (6). This process is repeated until $||U^{(t)} - U^{(t-1)}|| < \varepsilon$, that the ε is the termination criterion.

3.2.2 Kernel Fuzzy C-means Algorithm

The KFC method [23] [24] can map the initial data into a high-dimensional feature space. We use the kernels based on the Mercer theorem [25]. Adopting the kernel method can decrease computing time. In this algorithm X_i, $i = 1, 2, ..., N$ are the input data corresponding with the high-dimensional feature space data $\phi(X_c)$, $c = 1, 2, ..., C$. Based on Mercer theorem, $\phi(X_c)$ is a nonlinear mapping function.

Among the various kernel functions [26], polynomials $K(X, Y) = \phi(X) * \phi(Y) = (X * Y + b)^d$ and radial basis funcitons $K(X, Y) = \phi(X) * \phi(Y) = exp(-(X-Y)^2/2\sigma^2)$ are the most common. Using the kernel function according to Mercer theorem, we can map nonlinearly original input vectors into high-dimensional feature space. We do not use polynomials and other kernel functions because we determined that the radial-based kernel significantly outperforms other methods in our experiment.

Finally, we normalize the value and select the appropriate kernel function. After normalizing the vectors to a unit length, we use the following four distance methods to calculate the pairwise distances between two texts.

We employ four methods to compute the distance and obtain $KFC_{euclidean}$, KFC_{cosine}, $KFC_{jaccard}$, and $KFC_{pearson}$. The most appropriate parameter in our KFC experiment is $KFC_{euclidean}$. The distance between X_i and X_j and kernel function K can be defined as Equation (7).

$$d_{ij} = dist(\phi(X_i), \phi(X_j)) = \sqrt{\left\|\phi(X_i) - \phi(X_j)\right\|^2} \quad (7)$$

The distance d_{ij} can be expressed as:

$$d_{ij} = \sqrt{\phi(X_i)\phi(X_i) - 2\phi(X_i)\phi(X_j) + \phi(X_j)\phi(X_j)} = \sqrt{K(X_i,X_i) - 2K(X_j,X_j) + K(X_j,X_j)} \quad (8)$$

The degree of membership of all feature vectors in all clusters u_{ic} is computed using Equation (9) .

$$u_{ic} = \left(\frac{1}{d^2(X_c,V_i)}\right)^{\frac{1}{(m-1)}} \Big/ \sum_{i=1}^{C}\left(\frac{1}{d^2(X_c,V_i)}\right)^{\frac{1}{(m-1)}} \quad (9)$$

$K(X_c,\hat{V}_i)$ and $K(V_i,V_i)$ are updated as Equations (10) and (11).

$$K(X_c,\hat{V}_i) = \phi(X_c)\phi(\hat{V}_i) = \frac{\sum_{i=1}^{N}(u_{ic})^m K(X_i,X_c)}{\sum_{c=1}^{n}(u_{ic})^m} \quad (10)$$

$$K(\hat{V}_i,\hat{V}_i) = \phi(\hat{V}_i)\phi(\hat{V}_i) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(u_{ic})^m (u_{jc})^m K(X_i,X_j)}{\left(\sum_{c=1}^{n}(u_{ic})^m\right)^2} \quad (11)$$

Kernel Fuzzy C-means algorithm

Input:

Given a set of N data points $X = \{X_i\}_{i=1}^{N}$ and the number of clusters C and the termination criterion ε

Output:

The updated Matrix U

Steps:

1) Choose kernel function K and its parameters;
2) Initialize centroids $V_j = j = 1,2,...,C$;
3) Compute the degree of membership of all feature vectors in all the clusters using Equation (9),Where $d^2(X_c,X_i) = K(X_c,X_c) - 2K(X_c,V_i) + K(V_i,V_i)$;
4) Compute new kernel matrix $K(X_c,\hat{V}_i)$ and $K(\hat{V}_i,\hat{V}_i)$ using Equations (10) and (11);
5) Update $U^{(t)} = u_{ic}$ using Equation (9);
6) Repeat step 4 to step 6, until $\left\|U^{(t)} - U^{(t-1)}\right\| < \varepsilon$;
7) Rerurn $U^{(t)}$.

Figure 2. Kernel Fuzzy C-means algorithm. This algorithm is the pseudo code of the KFC algorithm. Input is the original data and parameters, and Output is the clustering result. The steps show the processes of the algorithm.

The pseudo code of the KFC algorithm is presented in Figure 2. Compared with the original FCM algorithm, the KFC method included the selection of Kernel functions. A few changes were carried out in Steps 4 and 5, as shown in Figure 2. The iteration stops and returns U when $||U^{(t)} - U^{(t-1)}|| < \varepsilon$, ε is termination criterion.

3.2.3 Comparison of FCM and KFC

We used the FCM and KFC two unsupervised clustering algorithms for words clustering. For each ambiguous word, 100 feature vectors represent the original contexts. We employed two algorithms to cluster the vectors. For each clustering algorithm, we set the number of the cluster

first. In the UMLS, each concept is assigned a concept unique identifier (CUI). Assigning a clustering number is a common issue that needs to be solved in unsupervised clustering methods. In our experiments, we followed the selection mechanism of the PPR [3] method. In the said paper, Agirre et al. identified the defined number of possible CUIs and instances for each ambiguous word. We considered the two numbers and adopted these numbers in our experiments directly.

Recently, many researchers focus on mapping the linear space to the corresponding nonlinear space using Mercer kernels. The kernel-based method is widely used for unsupervised learning. The Mark girolami et al. [27] adopted Mercer kernel for clustering and obtained a satisfactory result. Hsin-Chien Huang et al. [28] proposed a new type of kernel fuzzy clustering method using more than one kernel. The experiments showed that the clustering result of this algorithm was significantly improved.

We compared the FCM and KFC clustering algorithms on the NLM–WSD dataset. The key of the WSD results was considered as the clustering gold standard. We utilized the Rand index [29] algorithm to measure the statistical similarity between the clustering result and the gold standard. The evaluation results of the two algorithms are shown in Figure 3. The clustering results of the KFC algorithm are closer to the gold standard than the FCM algorithm results in most cases. Thus, we decided to use the KFC clustering algorithm in our experiments.

3.3 Sense Mapping

We use the KFC algorithm for clustering. In the unsupervised system, we measure the accuracy of our system based on the best alignment between the output clusters to the corpus-defined manual labeled gold standard.

Like all unsupervised systems, we could not ascertain the best match of the cluster and sense at first. Figure 4

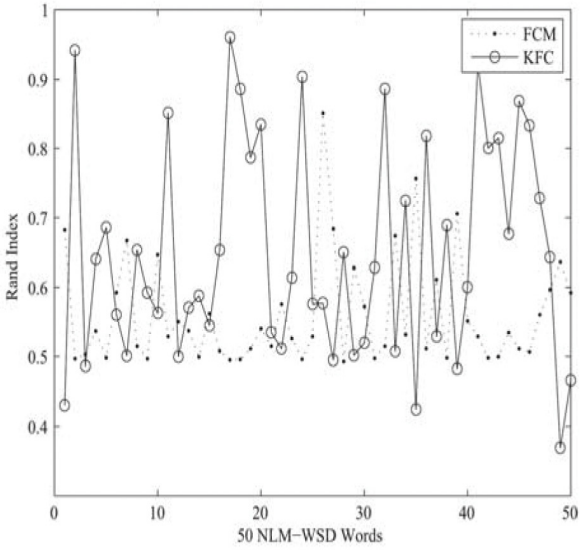

Figure 3. Rand Index Evaluation of FCM and KFC. The horizontal axis represents 50 ambiguous words. The vertical axis represents the Rand Index value of each algorithm. The KFC algorithm is indicated by a solid line, and the FCM algorithm is denoted by the dotted line

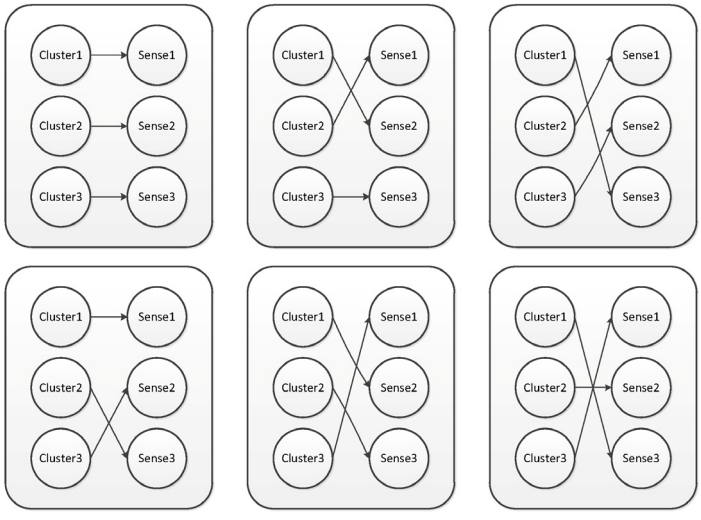

Figure 4. Possible Options of Sense Mapping. An example of all six possible matches of three clusters and three senses

presents an example of a combination of schemes of word mapping progress. In Figure 4, the unsupervised clustering result is three clusters. We aim to connect the clusters and the senses. The six possible connections of two sides are indicated Figure 4. In our experiment, we con sidered every possible match and obtained the best one as the final result for evaluation.

4. Experiments and Results

4.1 Experimental Settings

The NLM–WSD corpus includes 50 biomedical ambiguous terms (Table 1). Each ambiguous word has 100 contexts, which are extracted from the MEDLINE. The contexts include the title, id, abstract, and the keyword's position information. The instances of these terms were labeled by 11 annotators. Each instance is assigned to a certain sense. When no sense match exists in the instance, the instance is not labeled.

We first prepared datasets as the Leroy and Rindflesch (L&R) method. We chose 15 ambiguous terms from the

adjustment	association	blood- pressure	cold
condition	culture	degree	depression
determination	discharge	energy	evaluation
extraction	failure	fat	fit
fluid	frequency	ganglion	glucose
growth	Immuno-suppression	implantation	inhibition
japanese	lead	man	mole
mosaic	nutrition	pathology	pressure
radiation	reduction	repair	resistance
scale	secretion	sensitivity	sex
single	strains	support	surgery
transient	transport	ultrasound	variation
weight	white		

Table 1. WSD test collection list of words

NLM–WSD corpus, which were the same as (L&R) adopted in the Naïve Bayes system. The L&R subset is a balanced subset where the minority sense reaches at least 30% of the examples. Supervised methods require sufficient training data to determine the patterns of minority sense. For comparison, we used the most frequent sense (MFS) as the baseline. We also considered SENSATIONAL and FCM two unsupervised method for further comparison with our KFC method.

In the second part, the SENSATIONAL method evaluated the keywords divided into two categories: terms and acronyms. The terms part includes 13 ambiguous words where 12 terms come from the NLM–WSD corpus, whereas the other was the author's own extraction from PubMed abstracts. The acronyms part includes six acronyms extracted from PubMed. Obtaining exactly same the corpus as the SENSATIONAL method was not convenient. In our method, we used the 12 ambiguous terms from the NLM–WSD corpus , which were the same as the SENSATIONAL method. For the acronyms parts, we tested the ANA, BPD and BSA three acronyms, which were the intersection sets of the MSH–WSD corpus and SENSATIONAL method. Our acronym training data was not exactly the same as the SENSATIONAL method. Thus, this step was only a preliminary comparison of ambiguous acronyms. MFS and FCM are proposed as the baseline, which we also compared with that of the Sensecluster unsupervised system from Purandare and Pedersen [30].

In the third part of the experiment, the corpus includes all the 50 ambiguous terms from the NLM–WSD dataset. We compared the PPR system from Agirre et al. and the knowledge based systems CombSW and CombV from Jimeno-Yepes [20] with FCM and KFC system.

In our proposed method, the KFC algorithm is used for clustering. We measured the accuracy of our system by obtaining the best alignment between our output clusters and the manual labeled gold standard described in the sense mapping section.

The associated parameters of the KFC algorithm in our experiments are as follows: the kernel function is radial basis functions $exp(-(X-Y)^2/2\sigma^2)$, the fuzzy degree m = 1.08, the termination threshold $\varepsilon = 1e-5$, and the lower frequent words threshold is 3. The clustering algorithms stop when these algorithms reach the termination condition. After clustering, each ambiguous word is assigned

to a certain cluster. Mapping the right sense to each cluster is the final step of word sense disambiguation.

4.2 L&R subset results

Keywords (accuracy)	MFS	L&R	SENS-ATIO-NAL	FCM	KFC
adjustment	**0.62**	0.57	0.56	0.59	0.58
Blood-pressure	0.54	0.46	0.49	0.29	**0.55**
degree	0.63	0.68	0.72	0.63	**0.92**
evaluation	0.50	0.57	0.57	**0.65**	0.55
growth	0.63	0.62	**0.71**	0.44	0.64
Immuno-suppression	0.59	0.63	0.59	**0.70**	0.53
man	0.58	0.80	0.51	**0.82**	0.59
mosaic	0.52	0.66	**0.71**	0.70	0.57
nutrition	0.45	**0.48**	0.42	0.35	0.44
radiation	0.61	0.72	0.65	**0.81**	0.58
repair	0.52	0.81	0.80	**0.87**	0.77
scale	0.65	0.84	0.68	0.66	**0.95**
sensitivity	0.49	0.70	0.74	0.82	**0.86**
weight	0.47	**0.68**	0.53	0.49	0.57
white	0.49	**0.62**	0.52	0.34	0.49
Average accuracy	0.55	**0.66**	0.61	0.63	0.64

Table 2. Accuracy comparison on the L&R data set

The results of the L&R NLM data set are shown in Table 2. The best results for each term are highlighted in bold font. The precision of our KFC system outperforms that of the MFS baseline by an average of 9%. Our KFC system outperforms the SENSATIONAL system by an average of 3% in average precision. The KFC system is 1% higher than the FCM system when tested on the same parameters. The performance of the proposed system is not as good as the L&R system, which is a supervised system that utilizes the labeled training data as input. In fact, the KFC system exhibits a precision that is 2% lower than that of the L&R supervised system without employing manual labeled data.

4.3 Results on the subsets of SENSATIONAL

The unsupervised SENSATIONAL data group includes 12 terms and 3 acronyms. Table 3 shows that both the Senseclusters [30] and SENSATIONAL systems significantly improve the MFS baseline. The best results for each term are highlighted in bold font. The Senseclusters, which does not employ labeled data or manual resources,

Keywords (accuracy)	MFS	Sense-cluster	SENS-ATIO-NAL	FCM	KFC
cold	0.37	0.63	0.67	0.57	**0.82**
culture	0.52	0.55	**0.82**	0.61	0.68
discharge	0.66	0.90	**0.95**	0.72	0.92
fat	0.51	0.55	0.53	0.56	**0.88**
fluid	0.64	0.88	**0.99**	0.50	0.98
glucose	0.51	0.69	0.51	0.58	**0.91**
inhibition	0.5	0.55	0.54	0.52	**0.96**
mole	0.78	0.77	**0.96**	0.56	0.93
nutrition	0.39	0.50	**0.55**	0.35	0.44
pressure	0.52	0.89	0.86	0.58	**0.98**
sigle	0.50	0.87	**0.99**	0.62	0.96
transport	0.51	0.52	0.57	0.54	**0.97**
ANA	0.63	0.99	**1.00**	0.86	0.97
BPD	0.40	0.65	0.53	**0.99**	0.56
BSA	0.50	0.99	0.95	**1.00**	0.52
Average accuracy	0.53	0.73	0.76	0.64	**0.82**

Table 3. Accuracy comparison with unsupervised systems on terms and acronyms

achieves an improvement of 20% over the baseline precision of 53%. The SENSATIONAL exhibits an average precision of 76%. The KFC system shows an accuracy 6% higher than that of the SENSATIONAL system. In this subset, the SENSATIONAL system obtains 7 best results out of the 15 terms and acronyms, whereas our KFC system achieves 6 best results. The KFC system does not perform well in acronym disambiguation. We believe that the small corpus size is one of the reasons for this outcome.

4.4 Results on full NLM–WSD dataset

Table 4. Word by word accuracy results of full NLM–WSD dataset

Table 4 shows the results for 50 terms in the full NLM–WSD dataset. The column Word lists the ambiguous terms. The PPR column shows the word-by-word accuracy of Agirre's graph based Personalized PageRank method. The columns CombSW and CombV indicate the performance of Jimeno-Yepes's combination of four various unsupervised methods. The columns FCM and KFC are the two proposed unsupervised methods. The best results for each term are highlighted in bold font. The bottom row 'Best result rate' indicates the percentage of best results that each method obtains. The 'Average accuracy' is the final accuracy result of each column.

Keywords (accuracy)	PPR	Comb -SW	Comb V	FCM	KFC
adjustment	0.35	0.69	0.53	**0.59**	0.58
association	**1.00**	-	-	**1.00**	**1.00**
blood pressure	0.48	0.38	0.44	0.29	**0.55**
cold	0.28	0.39	0.79	0.57	**0.82**
condition	0.49	0.78	0.69	0.50	**0.88**
culture	0.77	**1.00**	0.55	0.61	0.68
degree	**0.94**	0.88	0.82	0.63	0.92
depression	0.94	0.97	**0.99**	0.53	0.91
determination	0.95	**0.96**	0.14	0.57	0.90
discharge	0.69	0.71	**0.96**	0.72	0.92
energy	0.28	0.46	0.54	0.62	**0.92**
evaluation	0.50	0.52	0.50	**0.65**	0.55
extraction	0.28	**0.98**	0.86	0.60	0.82
failure	0.72	0.86	**1.00**	0.48	0.76
fat	**0.96**	0.91	0.84	0.56	0.88
fit	0.11	0.89	**1.00**	0.94	**1.00**
fluid	0.92	0.49	0.35	0.50	**0.98**
frequency	0.99	0.63	0.81	0.51	**1.00**
ganglion	0.64	**0.88**	0.86	0.58	**0.88**
glucose	0.9	0.78	0.39	0.58	**0.91**
growth	0.37	0.55	**0.66**	0.44	0.64
Immuno-suppression	0.62	0.6	0.65	**0.70**	0.53
implantation	0.85	0.94	**0.97**	0.50	0.77
inhibition	0.22	**0.97**	0.83	0.52	0.96
japanese	0.65	0.63	**0.94**	0.56	0.92
lead	**0.93**	0.83	0.86	0.90	0.79
man	0.45	0.65	0.42	**0.82**	0.59
mole	0.27	**1.00**	**1.00**	0.56	0.93
mosaic	0.66	**0.85**	0.72	0.70	0.57
nutrition	0.33	**0.46**	0.43	0.35	0.44
pathology	0.28	0.76	**0.83**	0.44	0.77
pressure	**0.98**	0.64	0.88	0.58	**0.98**
radiation	0.53	0.77	0.77	**0.81**	0.58
reduction	0.55	**1.00**	0.82	0.82	0.82
repair	0.77	0.87	**0.88**	0.87	0.77
resistance	0.67	**1.00**	**1.00**	**1.00**	**1.00**
scale	0.85	0.69	0.66	0.66	**0.95**
secretion	0.99	0.58	**0.97**	0.54	0.81
sensitivity	0.28	**0.92**	0.73	0.82	0.86
sex	**0.85**	0.60	0.53	0.51	0.72
single	0.82	0.89	0.95	0.62	**0.96**
strains	0.97	**0.99**	0.96	0.53	0.96
support	0.8	**1.00**	0.90	0.70	0.80
surgery	**0.97**	0.43	0.96	0.62	0.8
transient	**0.99**	0.95	0.97	0.58	0.93
transport	0.69	**1.00**	0.98	0.54	0.97
ultrasound	0.83	0.81	0.83	0.57	0.84
variation	0.75	0.65	**0.86**	0.61	0.77
weight	0.57	0.66	**0.68**	0.49	0.57
white	0.63	0.57	0.58	**0.72**	0.60
Best result rate	0.18	0.31	0.24	0.14	**0.32**
Average accuracy	0.67	0.76	0.76	0.62	**0.81**

Table 4. Word by word accuracy results of full NLM–WSD dataset

The PPR algorithm performs various terms with results ranging from 11.1% to 99%. The overall performance of the PPR algorithm depends on the graph situations. Unfavorable graph situations adversely affect the average accuracy. Although the PPR obtains 18 best results in 50 terms, the average accuracy reaches 67%, which is slightly lower than that of many other methods. The Jimeno-Yepes's CombSW and CombV combine four knowledge-based methods. Their works rely on the UMLS Metathesaurus ontology. The first method compares the overlap of the context using definition, synonyms, and related terms. The second method uses a graph-based method on the Metathesaurus network to perform unsupervised WSD. The third approach employs PubMed to collect extra training data and adopts the WSD work by Naïve Bayes. The last approach utilizes the semantic types with the Metathesaurus to perform WSD. The CombSW and CombV are two combination strategies, comprising weighted linear combination and voting combination. The performances of these approaches on most of the terms are outstanding, and the average accuracies of these methods reach 76%.

In the group of our methods, the FCM stands for fuzzy C-means and the KFC combines the kernel function with the FCM clustering. The FCM shows a lower accuracy than the other least-performing methods. The average accuracy of KFC is 81%, which is significantly higher than all the other methods. The best result rate of KFC is 32%, which is slightly higher than 31% of the CombSW. The results demonstrate that using the kernel function has a significant positive effect in the WSD system. In this task, the KFC method clearly outperforms other WSD methods, as indicated in Table 4.

We also compare the proposed approach with the Stevenson's supervised system, which we did not list in the Table 4. This system combines three supervised methods and achieves an accuracy of 0.88. The accuracy of our system at the same dataset is 0.81, which is close to the state-of-art supervised system.

5. Conclusion

We compared several methods for word sense disambiguation in the biomedical domain. We determined that the kernel fuzzy c-means clustering method outperforms other unsupervised methods. The average accuracy of KFC method is close to the state-of-art supervised methods. Based on our analysis, the reasons are as follows:

The NLM–WSD corpus has 50 ambiguous terms, and each term has 100 contexts. The corpus data is unbalanced because some senses obtained very few contexts out of 100 contexts. In this situation, training data is insufficient for some classical supervised methods. Imbalanced data in the biomedical texts is very common, which must be addressed in the supervised methods. Unsupervised methods discriminate the senses first and map the senses based on similarity or external resources. Thus, these approaches usually operate well on imbalance data.

The KFC method maps the features on high-dimensional space. Distinguishing the similar senses is more convenient. The sense mapping strategy is effective for imbalanced data. Thus, the final results performed well in terms of average accuracy.

In this paper, we use single kernel method for NLM–WSD task. We will explore multiple kernels methods in further research to enhance the current research. We will also focus on acronym disambiguation in the biomedical domain.

Acknowledgement

This work was Supported by the Fundamental Research Funds for the Central Universities, South-Central University for Nationalities (CZQ14012).

References

[1] Li, L., Guo, R., Jiang, Z., Huang, D. (2014). Improving Kernel-based protein-protein interaction extraction by unsupervised word representation. *In:* Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'2014*),* p. 379-384, IEEE, Belfast, Nov. 2014.

[2] Duan, W., Song, M., Yates, A. (2009). Fast max-margin clustering for unsupervised word sense disambiguation in biomedical texts. *BMC bioinformatics*, 10(Suppl 3), S4.

[3] Agirre, E., Soroa, A.,Stevenson, M. (2010). Graph-based Word Sense Disambiguation of biomedical documents. *Bioinformatics*, 26 (22) 2889-2896.

[4] Ginter, F., Boberg, J., Järvinen, J. (2004). New techniques for disambiguation in natural language and their application to biological text. *The Journal of Machine Learning Research*, (5) 605-621.

[5] Liu, H., Teller, V., Friedman, C. (2004). A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. *Journal of the American Medical Informatics Association*, 11 (4) 320-331.

[6] Leroy, G., Rindflesch, T. C. (2004). Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naive Bayes classifier. *Medinfo*, 11 (1) 381-385.

[7] Joshi, M., Pedersen, T. et al. (2006). Kernel Methods for Word Sense Disambiguation and Acronym Expansion. *American Association for Artificial Intelligence (AAAI)*, (6) 1879-1880.

[8] Savova, G. K., Coden, A. R. et al. (2008). Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41 (6) 1088-1100.

[9] Stevenson, M., Guo, Y. (2010). The Effect of Ambiguity on the Automated Acquisition of WSD Examples. *In: Proc. of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'2010),* p. 353-356. NAACL, Los Angeles, Jun.2010.

[10] Al-Mubaid, H., Gungu, S. (2012). A learning-based approach for biomedical word sense disambiguation. *The Scientific World Journal*, (2012) 1-8.

[11] Savova, G., Pedersen, T., et al. (2005). Resolving Ambiguities in Biomedical Text with Unsupervised Clustering Approaches. http://www.d.umn.edu/~kulka020/ResearchReport_2005-80.pdf.

[12] Schijvenaars, B., Mons, B., Weeber, M., et al. (2005). Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, 6 (1) 149.

[13] Pahikkala, T., Pyysalo, S., Ginter, F., et al. (2005). Kernels incorporating word positional information in natural language disambiguation tasks. *In: Proc. of the Eighteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2005)*, p. 442-448. AAAI, Clearwater Beach, May. 2005.

[14] Liu, H., Teller, V., Friedman, C. (2004). A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. *Journal of the American Medical Informatics Association*, 11 (4) 320-331.

[15] Gaudan, S., Kirsch, H., Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21 (18) 3658-3664.

[16] Humphrey, S. M., Rogers, W. J. et al. (2006). Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57 (1) 96-113.

[17] Navigli, R., Lapata, M. (2007). Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. *In: Proc. of the 20th international joint conference on Artifical intelligence(IJCAI'2007)*, p. 1683-1688. AAAI, Hyderabad, Jan. 2007.

[18] Sinha, R., Mihalcea, R. (2007). Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. *In: Proc. of the International Conference on Semantic Computing(ICSC'2007),* p. 363-369. IEEE, Irvine, Sep. 2007.

[19] Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I. (2007). Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. *In: Proc. of the 20th International Joint Conference on Artifical intelligence(IJCAI'2007)*, p. 1725-1730. AAAI, Hyderabad, Jan. 2007.

[20] Jimeno-Yepes, A. J., Aronson, A. R. (2011). Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC bioinformatics*, 11 (1) 569.

[21] Jimeno-Yepes, A. J., McInnes, B. T., Aronson, A. R. (2011). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12 (1) 223.

[22] Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York and London: Plenum Press.

[23] Zhang, D. Q., Chen, S. C. (2003). Clustering incomplete data using kernel-based fuzzy c-means algorithm. *Neural Processing Letters*, 18 (3) 155-162.

[24] Wu, Z., Xie, W., Yu, J. (2003). Fuzzy c-means clustering algorithm based on kernel method. Computational Intelligence and Multimedia Applications. *In: Proc. of the Computational Intelligence and Multimedia Applications Proceedings (ICCIMA'2003),* p. 49-54. IEEE, Xi'an, Sept. 2003.

[25] Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A,* 209 (1909) 415-446.

[26] Burges, C. J. (1999). Geometry and invariance in kernel based methods. http://dl.acm.org/citation.cfm?id=299100.

[27] Girolami, M. (2002). Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks,* 13 (3) 780-784.

[28] Huang, H. C., Chuang, Y. Y., Chen, C. S. (2012). Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems,* 20 (1) 120-134.

[29] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* 66 (336) 846-850.

[30] Purandare, A., Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. *In: Proc. of the Conference on Computational Natural Language Learning(CoNLL'2004)*, p. 41-48. ACL, Boston, May. 2004.