Cross-Lingual and Supervised Learning Approach for Indonesian Word Sense Disambiguation Task

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and **Mirna Adriani**
Faculty of Computer Science, Universitas Indonesia
Depok 16424, West Java, Indonesia
rahmad.mahendra@cs.ui.ac.id, {heninggar, haryo97}@gmail.com

Abstract

Ambiguity is a problem we frequently face in Natural Language Processing. Word Sense Disambiguation (WSD) is a task to determine the correct sense of an ambiguous word. However, research in WSD for Indonesian is still rare to find. The availability of English-Indonesian parallel corpora and WordNet for both languages can be used as training data for WSD by applying Cross-Lingual WSD method. This training data is used as an input to build a model using supervised machine learning algorithms. Our research also examines the use of Word Embedding features to build the WSD model.

1 Introduction

One of the biggest challenges in Natural Language Processing (NLP) is ambiguity. Ambiguity exists when there are many alternatives of linguistic structures that can be composed for an input language. Some words can have more than one meaning (word sense). For example the word "kali" in Indonesian can posses two senses, i.e. river and frequency (as described in Table 1)

Word Sense Disambiguation (WSD) is a task to determine the correct sense of a polysemous word. Even though it becomes a fundamental task in NLP, research on WSD for Indonesian language has not attracted many interests. To our knowledge, the only published work was Uliniansyah and Ishizaki (2005). Uliniansyah and Ishizaki applied the corpus-based approach using Naive Bayes as the classifier. The training data was collected from news websites and manually annotated. The words in training data were processed using the morphological analysis to obtain lemma. The features being used were some words around the target word (including the words before and after the target word), the nearest verb from the target word, the transitive verb around the target word, and the document context. Unfortunately, neither the model nor the corpus from this research is made publicly available.

This paper reports our study on WSD task for Indonesian using the combination of the cross-lingual and supervised learning approach. Training data is automatically acquired using Cross-Lingual WSD (CLWSD) approach by utilizing WordNet and parallel corpus. Then, the monolingual WSD model is built from training data and it is used to assign the correct sense to any previously unseen word in a new context.

2 Related Work

WSD task is undertaken in two main steps, namely listing all possible senses for a word and determining the right sense given its context (Ide and Véronis, 1998). To list possible senses, we can use dictionaries, knowledge resources (e.g. thesaurus, WordNet), and transfer directory (e.g. translation from other language). To determine the right sense, we can use the information from the context where the word is used, and also external knowledge resource such as dictionary or encyclopedia.

Among various approaches to WSD, supervised learning approach is the most successful one to date. The supervised WSD uses machine learning techniques for inducing a classifier from sense-annotated data sets. Training data used to learn the classifier contains a set of examples in which each occurrence of an ambiguous word has been annotated with the correct sense according to existing sense inventory. (Navigli, 2009)

Despite of its success, the supervised learning approach has a drawback of requiring manually sense-tagged data. Manually labeling data for training set is costly and time-consumptive. As an alternative, the sense labeling can be done automatically by utilizing existing resources. Cross

Sentence in Indonesian	English translation	Meaning of "kali"
Saya makan dua **kali** pagi ini	I ate twice this morning	frequency
Rumah saya di dekat **kali**	My house is near the river	river

Table 1: Word Ambiguity Example in Bahasa Indonesia

lingual approach is able to disambiguate word sense based on the evidence from the translation information. The rationale behind this approach is that a different sense of a word typically has different translations in other languages. The plausible translations of a word in context restrict the number of its possible senses. Cross-Lingual Word Sense Disambiguation (CLWSD) aims to automatically disambiguate a text in one language by exploiting its differences to other language(s) in a parallel corpus.

Before being a dedicated task in SemEval-2013 (Lefever and Hoste, 2010), CLWSD has been explored in several works. Brown et.al. (1991) proposed an unsupervised approach for WSD. The word alignment was performed on a parallel corpus, and then the most appropriate translation was determined for a target word based on a set of contextual features.

Ide et.al (2002) conducted an experiment using translation equivalents derived from parallel corpus to determine the sense distinctions that can be used for automatic sense-tagging and other disambiguation tasks. They found that sense distinctions derived from cross-lingual information are at least as reliable as those made by human annotators. In their study on seven languages (English, Romanian, Slovene, Czech, Bulgarian, Estonian, and Hungarian), Ide et.al exploited EuroWordNet as a knowledge source.

Sense intersection, an approach described in Gliozzo et al. (2005) and Bonansinga and Bond (2016), inspires CLWSD process in our study. Gliozzo et.al. proposed an unsupervised WSD technique to automatically acquire sense tagged data that exploited the polysemic differential between two languages using aligned corpora and multilingual lexical databases. An aligned multilingual lexical resource (e.g. MultiWordNet) allowed them to disambiguate aligned words in both languages by simply intersecting their senses. Bond and Bonansinga (2015) then applied the sense intersection approach in multilingual settings. Bonansinga and Bond (2016) considered four languages, e.g. English, Italian, Romanian, and Japanese in their experiment to reduce more ambiguity.

For the supervised learning approach to WSD tasks, the common features are the surrounding words of target word, POS tags of the surrounding words, and local collocation. Current studies (Taghipour and Ng, 2015) (Iacobacci et al., 2016) examined the potential use of Word Embedding as a feature for the sense classification task. Iacobacci et.al. (2016) described four different strategies to use Word Embedding, e.g. concatenation, average, fractional decay, and exponential decay.

3 CLWSD for Building Indonesian WSD Training Data

We utilize CLWSD using parallel corpus and WordNet to acquire WSD training data. The model is then learned from the training data to disambiguate the word sense in testing data. Our CLWSD approach is illustrated in Figure 1.

The input for CLWSD process is English-Indonesian parallel corpus. The corpus used in our experiment is Identic++, which is Identic corpus (Larasati, 2012) that has been extended by adding the instances of English-Indonesia parallel sentences from movie subtitles. In addition to the parallel corpus, we harnessed lexical database, namely Princeton WordNet (Miller, 1995) and WordNet Bahasa (Noor et al., 2011).

CLWSD process consists of several steps:

1. Align the words in the parallel corpus using GIZA++ (Och and Ney, 2003) to obtain translation pairs.

2. Assign the sense label to the words by using sense-ID from WordNet. English words are labeled with the sense-ID from Princeton WordNet, and Indonesian words are labeled with the sense-ID from Indonesian WordNet. There may be more than one possible sense-ID for a single word. To disambiguate the word sense, we find the intersection between English and Indonesian sense inventory of the words in the translation pairs. Since our

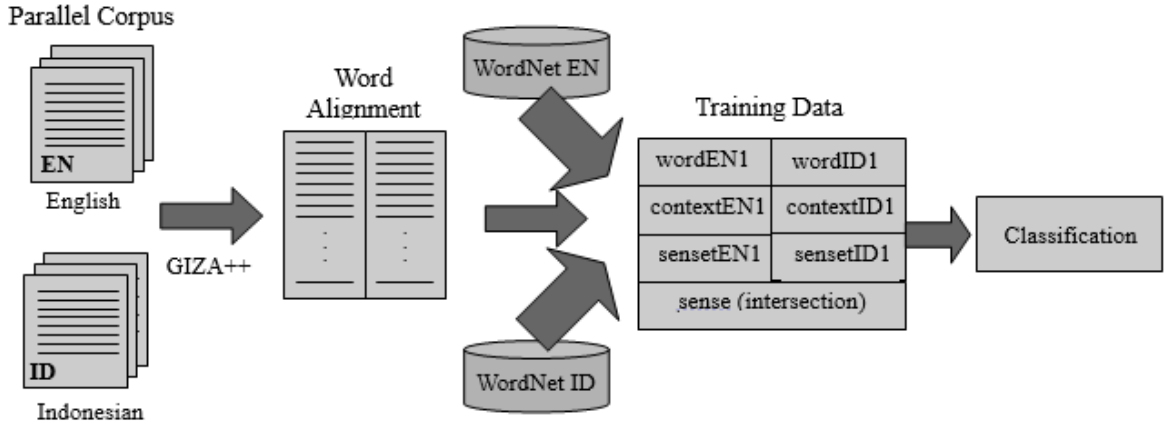

Figure 1: CLWSD Methodology

study aims to obtain the training data with high precision, we consider only the word pair instances that have exactly one intersected sense-ID.

3. Extract the content words surrounding each word to be disambiguated in the corpus.

An example is given to illustrate the process. Consider the following entry of the parallel corpus.

> "EN: `She reads page 50 of that book`".
> "ID: `Dia membaca halaman 50 dari buku itu`".

That sentence pair (along with the rests in the corpus) is processed with GIZA++. Word "`page`" is aligned with "`halaman`". The sense labels for the words "`page`" and "`halaman`" are listed below.

> `page` → **06256697**, 11220149, 10391416, 10391248, 10391086
> `halaman` → 00193486, 00227165, 00754560, **06256697**

Among many senses corresponding to each English and Indonesian words, there is one intersected sense. Therefore, the word "`halaman`" in the sentence "`dia sedang membaca halaman 50 dari buku itu`" is labelled with the sense-ID 06256697. Moreover, the content words in this sentence include "`dia`", "`sedang`", "`membaca`", "`dari`", "`buku`", and "`itu`".

Word	Translation	Number of Instances
alam	nature	251
	universe	225
atas	above	546
	top	441
kayu	timber	51
	wooden	49
anggur	wine	272
	grape	21
perdana	prime	302
	premier	22
dasar	primary	225
	underlying	11
All samples		2,416

Table 2: Sample Words for Monolingual WSD

We have retrieved 352,816 pairs of aligned words between Indonesian and English from the Identic++ corpus. Among of them, 4,237 Indonesian words are polysemous. The rest of words may have only one sense (not ambigue) or no corresponding sense found in WordNet towards them. Finally, 752 different words can be disambiguated using sense intersection approach.

4 Supervised Learning for Indonesian Language WSD

The sense-tagged words acquired in CLWSD process are used to train classifier. The classifier induced the model for Indonesian WSD. For evaluation of the supervised learning approach, we performed monolingual lexical sample WSD task. We tested sample of 2,416 sentences that contain

Word	Baseline	NB	MLP	RF	SVM	XGB
alam	36.41	62.45	94.54	94.75	95.17	**96.01**
atas	39.41	69.79	71.95	71.16	71.69	**72.21**
kayu	34.45	69.98	66.52	71.98	70.71	**73.98**
anggur	89.38	89.17	**91.81**	89.86	90.92	89.40
perdana	90.03	89.12	**93.77**	91.23	91.64	91.04
dasar	93.06	92.42	**95.90**	93.06	94.29	94.29
Average	63.79	78.82	85.75	85.34	85.74	**86.16**

Table 3: F1 Score of Baseline vs Machine Learning Models using BoW Features

one of 6 target words. Each of these target words has two possible senses. Sample Indonesian words for monolingual WSD experiment are listed in Table 2.

We ran the experiment in 10-fold cross validation setting. We built the model using five different supervised machine learning algorithms, namely Naive Bayes, Multi Layer Perceptron (MLP), Random Forest (Breiman, 2001), Support Vector Machine (SVM) (Boser et al., 1992), and XGBoost (Chen and Guestrin, 2016). For the baseline evaluation, we assign the most frequent sense label to each instance.

Using the content words as bag-of-words (BoW) representation, any machine learning models tested in our experiment outperformed the baseline evaluation. All machine learning models but Naive Bayes obtained the average F-1 score >85%. XGBoost model achieved the best average F-1 score, that is 86.16%. On other hand, MLP performed better compared to other models to disambiguate the words with imbalanced sense label distribution (e.g. "`anggur`", "`perdana`", and "`dasar`"). Complete evaluation of the baseline and machine learning methods is presented in Table 3.

4.1 POS and Word Embedding as Features

A word in the different parts of speech (POS) has the different sense. A word used in the different senses is likely to have the different set of POSs around it. So, the POS information of content words can be potential cue to determine the word sense.

To obtain the POS feature, we used Indonesian POSTag model from Rashel et.al (2014). In general, incorporating the POS into the bag-of-words features improve WSD performance in our experiment. Average F-1 scores of SVM and MLP models increase, but there is a slight decrease in F-1 score of XGBoost model.

Beside that, we conducted other experiments using the Word Embedding features. We transformed each word in the sentence into continous-space vector representation using skip gram model pre-trained by Word2Vec (Mikolov et al., 2013). We considered two different strategies to incorporate the Word Embedding in monolingual Indonesian WSD task. First, the vectors of the content words are concatenated into a larger vector that has a size equal to the aggregated dimension of all the individual embeddings (**concat**). Second, the vectors of the content words are summed up and the resultant vector is divided by number of content words (**avg**).

Sense classification using the Word Embedding features produced promising result. MLP and XGBoost model that make use of the Word Embedding on the basis of average strategy reach the F-1 score respectively 86.80% and 86.34%. These scores are higher than the best result achieved by same models using the traditional bag-of-words only. The experimental result related to use of the features in lexical sample WSD task is reported in Table 4.

4.2 Effect of Stemming and Stopword Removal to BoW Features

Stemming is a common technique used in information retrieval to eliminate the morphology variations to obtain the basic form of a word. On the other hand, stopword removal is the process of removing common words that are often used in many sentences, e.g. "`and`", "`or`", "`is`". We had a hypothesis that the stemming and stopword removal can affect the WSD system performance. Stemming is used in order that the words with different morphological forms can be counted as the same content words. In addition, stopwords removal is used to prevent the matrix representing

Word	BoW			BoW+POS			WE (concat)			WE (avg)		
	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB
alam	95.17	94.54	96.01	95.60	94.13	**96.22**	89.10	88.26	85.53	87.42	89.94	92.87
atas	71.69	71.95	72.21	73.29	**74.72**	74.38	63.72	67.38	68.37	67.20	69.81	70.10
kayu	70.71	66.52	73.97	73.72	67.91	68.36	71.25	78.21	74.23	82.12	**82.16**	77.21
anggur	90.92	91.81	89.40	90.42	92.06	88.68	87.72	89.01	89.79	91.84	**92.23**	91.06
perdana	91.64	**93.77**	91.03	93.54	93.54	92.36	90.11	89.04	90.83	89.95	91.03	91.12
dasar	94.28	**95.90**	94.28	93.06	92.85	94.98	93.77	92.64	92.85	95.64	95.65	95.67
Average	85.74	85.75	86.16	86.61	85.87	85.83	82.61	84.09	83.60	85.70	**86.80**	86.34

Table 4: WSD Experiment Using POS and Embedding Features

Word	BoW			stem			no stopword			stem & no stopword		
	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB
alam	95.17	94.54	96.01	95.80	93.49	95.80	95.38	92.85	96.22	96.01	93.48	**96.43**
atas	71.69	71.95	72.21	71.00	**74.18**	72.06	69.31	70.77	63.89	70.82	72.09	67.22
kayu	70.71	66.52	73.98	78.79	70.96	67.97	75.59	61.46	70.45	**82.89**	65.89	70.96
anggur	90.92	**91.81**	89.40	91.41	90.17	89.60	90.26	91.56	88.87	90.60	91.31	89.38
perdana	91.64	93.77	91.04	93.01	93.03	91.04	92.80	**94.43**	92.21	92.80	93.01	91.64
dasar	94.29	95.90	94.29	95.00	**96.25**	94.59	93.06	94.90	94.07	94.04	93.06	94.67
Average	85.74	85.75	86.16	87.50	86.35	85.18	86.07	84.33	84.29	**87.86**	84.81	85.05

Table 5: Effect of Stemming and Stopword Removal to WSD Model

content words becomes too sparse, as well as to remove unimportant words from the content words.

We used the Indonesian stemmer (Adriani et al., 2007) to derive the stem of content word, while stopword removal was conducted using dictionary of Indonesian stopwords (Tala, 2003).

The effect of stemming in this study is increasing the F1-score (for SVM and MLP model). The initial F1-score of SVM model using the bag-of-words feature is 85.74% and after the stemming the F1-score becomes 87.50%. The words, that were previously considered different because of the morphological variations, are counted as the same words after the stemming, so two sentences that were considered unlike now become similar. On the other hand, the effect of stopwords removal is not as good as stemming. MLP and XGBoost models have decreased the F1-scores when the stopwords are excluded from the bag-of-words feature. We argue that the stopwords list may still contain the words that are discriminative enough to explain the context of the sentence.

5 Summary

In our study, CLWSD has been implemented to provide the training data and then the model based on the training data is built by the classifier to perform monolingual Indonesian WSD. We took advantage of existing of the parallel corpus and WordNet to obtain the sense-labeled words by a cross lingual approach. We retrieved all possible senses for the translation pairs and then found the intersection between senses from English and Indonesian language. The data acquired by CLWSD process is released at `https://github.com/rmahendra/Indonesian-WSD`. We ran several experiments on monolingual WSD task and concluded that any supervised machine learning model outperforms the baseline method. Moreover, we found that the use of embedding vector can produce better F1-score of sense classification than the use of the traditional bag-of-words features.

The study still has rooms for improvement. We need to test our methodology in larger corpus and involve more target words for experiment. Detail evaluation of CLWSD to produce Indonesian training data can be more explored. On the other hand, it is interesting to check how sensitive our proposed approach works when considering the semantic difference between senses.

Acknowledgments

The authors gratefully acknowledge the support of the PITTA UI Grant Contract No. 410/UN2.R3.1/HKP.05.00/2017

References

Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi, and Hugh E. Williams. 2007. Stemming indonesian: A confix-stripping approach. *Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33, December.

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proceedings of the Eighth Global WordNet Conference*, pages 44–49.

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In: Second Italian Conference on Computational Linguistics CLiC-it.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, October.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 264–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for wsd. In *In: Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, vol 3406*, CICLing 2005, pages 242–245, Berlin, Heidelberg. Springer.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.

Nancy Ide and Jean Véronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):2–40.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8*, WSD '02, pages 61–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Septina Dian Larasati. 2012. Identic corpus: Morphologically enriched indonesian english parallel corpus. In *Proceedings of LREC*.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communication of ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10:1–10:69, February.

Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.

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

Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an indonesian rule-based part-of-speech tagger. In *Proceedings of 2014 International Conference on Asian Language Processing (IALP)*. IEEE.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 314–323.

Fadillah Z Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia. Master's thesis, University of Amsterdam, the Netherlands.

Mohammad Teduh Uliniansyaht and Shun Ishizaki. 2005. A word sense disambiguation system using modified naive bayesian algorithms for indonesian language. *Natural Language Processing*, 12(1):33–50.