Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation

Akihiro Tamura and **Taro Watanabe** and **Eiichiro Sumita**
Multilingual Translation Laboratory, MASTAR Project
National Institute of Information and Communications Technology
3-5 Hikaridai, Keihanna Science City, Kyoto, 619-0289, JAPAN
{akihiro.tamura,taro.watanabe,eiichiro.sumita}@nict.go.jp

Abstract

This paper proposes a novel method for lexicon extraction that extracts translation pairs from comparable corpora by using graph-based label propagation. In previous work, it was established that performance drastically decreases when the coverage of a seed lexicon is small. We resolve this problem by utilizing indirect relations with the bilingual seeds together with direct relations, in which each word is represented by a distribution of translated seeds. The seed distributions are propagated over a graph representing relations among words, and translation pairs are extracted by identifying word pairs with a high similarity in the seed distributions. We propose two types of the graphs: a co-occurrence graph, representing co-occurrence relations between words, and a similarity graph, representing context similarities between words. Evaluations using English and Japanese patent comparable corpora show that our proposed graph propagation method outperforms conventional methods. Further, the similarity graph achieved improved performance by clustering synonyms into the same translation.

1 Introduction

Bilingual lexicons are important resources for bilingual tasks such as machine translation (MT) and cross-language information retrieval (CLIR). Therefore, the automatic building of bilingual lexicons from corpora is one of the issues that have attracted many researchers. As a solution, a number of previous works proposed extracting bilingual lexicons from comparable corpora, in which documents were not direct translations but shared a topic or domain[1]. The use of comparable corpora is motivated by the fact that large parallel corpora are only available for a few language pairs and for limited domains.

Most of the previous methods are based on assumption (I), that a word and its translation tend to appear in similar contexts across languages (Rapp, 1999). Based on this assumption, many methods calculate word similarity using context and then extract word translation pairs with a high-context similarity. We call these methods context-similarity-based methods. The context similarities are usually computed using a seed bilingual lexicon (e.g. a general bilingual dictionary) by mapping contexts expressed in two different languages into the same space. In the mapping, information not represented by the seed lexicon is discarded. Therefore, the context-similarity-based methods could not find accurate translation pairs if using a small seed lexicon.

Some of the previous methods tried to alleviate the problem of the limited seed lexicon size (Koehn and Knight, 2002; Morin and Prochasson, 2011; Hazem et al., 2011), while others did not require any seed lexicon (Rapp, 1995; Fung, 1995; Haghighi et al., 2008; Ismail and Manandhar, 2010; Daumé III and Jagarlamudi, 2011). However, they suffer the problems of high computational cost (Rapp, 1995), sensitivity to parameters (Hazem et al., 2011), low accuracy (Fung, 1995; Ismail and Manandhar, 2010), and ineffectiveness for language pairs with

[1]Although Vulić et al. (2011) regarded document-aligned texts such as texts on Wikipedia as comparable corpora, we do not limit comparable corpora to these kinds of texts.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 24–36, Jeju Island, Korea, 12–14 July 2012.

different types of characters (Koehn and Knight, 2002; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011).

In face of the above problems, we propose a novel method that uses a graph-based label propagation technique (Zhu and Ghahramani, 2002). The proposed method is based on assumption (II), which is derived by recursively applying assumption (I) to the "contexts": a word and its translation tend to have similar co-occurrence (direct and indirect) relations with all bilingual seeds across languages.

Based on assumption (II), we propose a three-step approach: (1) constructing a graph for each language with each edge indicating a direct co-occurrence relation, (2) representing every word as a seed translation distribution by iteratively propagating translated seeds in each graph, (3) finding two words in different languages with a high similarity with respect to the seed distributions. By propagating all the seeds on the graph, indirect co-occurrence relations are also considered when computing bilingual relations, which have been neglected in previous methods. In addition to the co-occurrence-based graph construction, we propose a similarity graph, which also takes into account context similarities between words.

The main contributions of this paper are as follows:

- We propose a bilingual lexicon extraction method that captures co-occurrence relations with all the seeds, including indirect relations, using graph-based label propagation. In our experiments, we confirm that the proposed method outperforms conventional context-similarity-based methods (Rapp, 1999; Andrade et al., 2010), and works well even if the coverage of a seed lexicon is low.
- We propose a similarity graph which represents context similarities between words. In our experiments, we confirm that a similarity graph is more effective than a co-occurrence-based graph.

2 Context-Similarity-based Extraction Method

The bilingual lexicon extraction from comparable corpora was pioneered in (Rapp, 1995; Fung, 1995). The popular similarity-based methods consist of the following steps: modeling contexts, calculating context similarities, and finding translation pairs.

Step 1. Modeling contexts: The context of each word is generally modeled by a vector where each dimension corresponds to a context word and each dimension has a value indicating occurrence correlation. Various definitions for the context have been used: distance-based context (e.g. in a sentence (Laroche and Langlais, 2010), in a paragraph (Fung and McKeown, 1997), in a predefined window (Rapp, 1999; Andrade et al., 2010)), and syntactic-based context (e.g. predecessors and successors in dependency trees (Garera et al., 2009), certain dependency position (Otero and Campos, 2008)). Some treated context words equally regardless of their positions (Fung and Yee, 1998), while others treated the words separately for each position (Rapp, 1999). Various correlation measures have been used: log-likelihood ratio (Rapp, 1999; Chiao and Zweigenbaum, 2002), tf-idf (Fung and Yee, 1998), pointwise mutual information (PMI) (Andrade et al., 2010), context heterogeneity (Fung, 1995), etc.

Shao and Ng (2004) represented contexts using language models. Andrade et al. (2010) used a set of words with a positive association as a context. Andrade et al. (2011a) used dependency relations instead of context words. Ismail and Manandhar (2010) used only in-domain words in contexts. Pekar et al. (2006) constructed smoothed context vectors for rare words. Laws et al. (2010) used graphs in which vertices correspond to words and edges indicate three types of syntactic relations such as adjectival modification.

Step 2. Calculating context similarities: The contexts which are expressed in two different languages are mapped into the same space. Previous methods generally use a seed bilingual lexicon for this mapping. After that, similarities are calculated based on the mapped context vectors using various measures: city-block metric (Rapp, 1999), cosine similarity (Fung and Yee, 1998), weighted jaccard index (Hazem et al., 2011), Jensen-Shannon divergence (Pekar et al., 2006), the number of overlapping context words (Andrade et al., 2010), SimRank (Laws et al., 2010), euclidean distance (Fung, 1995), etc.

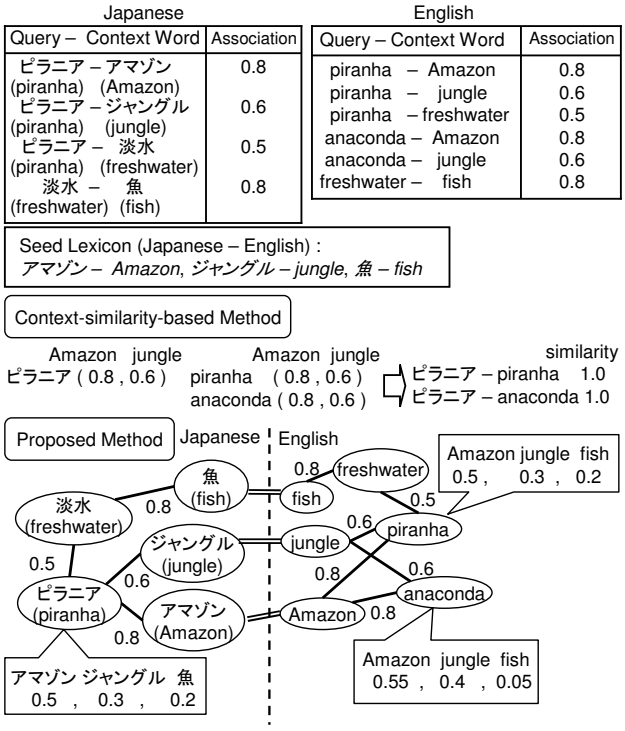

Figure 1: An Example of a Previous Method and our Proposed Method

Andrade et al. (2011b) performed a linear transformation of context vectors in accordance with the notion that importance varies by context positions. Gaussier et al. (2004) mapped context vectors via latent classes to capture synonymy and polysemy in a seed lexicon. Fišer et al. (2011) and Kaji (2005) calculated 2-way similarities.

Step 3. Finding translation pairs: A pair of words is treated as a translation pair when their context similarity is high. Various clues have been considered when computing the similarities: concept class information obtained from a multilingual thesaurus (Déjean et al., 2002), co-occurrence models generated from aligned documents (Prochasson and Fung, 2011), and transliteration information (Shao and Ng, 2004).

2.1 Problems from Previous Works

Most of previous methods used a seed bilingual lexicon for mapping modeled contexts in two different languages into the same space. The mapping heavily relies on the entries in a given bilingual lexicon. Therefore, if the coverage of the seed lexicon is low, the context vectors become sparser and its discriminative capability becomes lower, leading to extraction of incorrect translation equivalents.

Consider the example in Figure 1, where a context-similarity-based method and our proposed method find translation equivalents of the Japanese word "ピラニア (piranha)". There are three context words for the query. However, the information on co-occurrence with "淡水 (freshwater)" disappears after the context vector is mapped, because the seed lexicon does not include "淡水 (freshwater)". The same thing happens with the English word "piranha". As a result, the pair of "ピラニア (piranha)" and "anaconda" could be wrongly identified as a translation pair.

Some previous work focused on the problem of seed lexicon limitation. Morin and Prochasson (2011) complemented the seed lexicon with bilingual lexicon extracted from parallel sentences. Koehn and Knight (2002) used identically-spelled words in two languages as a seed lexicon. However, the method is not applicable for language pairs with different types of characters such as English and Japanese. Hazem et al. (2011) exploited *k*-nearest words for a query, which is very sensitive to the parameter *k*.

Some previous work did not require any seed lexicon. Rapp (1995) proposed a computationally demanding matrix permutation method which maximizes a similarity between co-occurrence matrices in two languages. Ismail and Manandhar (2010) introduced a similarity measure between two words in different languages without requiring any seed lexicon. Fung (1995) used context heterogeneity vectors where each dimension is independent on language types. However, their performances are worse than those of conventional methods using a small seed lexicon. Haghighi et al. (2008) and Daumé III and Jagarlamudi (2011) proposed a generative model based on probabilistic canonical correlation analysis, where words are represented by context features and orthographic features[2]. However, their experiments showed that orthographic features to be important for effectiveness, which means low per-

[2]In Haghighi et al. (2008) and Daumé III and Jagarlamudi (2011), indirect relations with seeds are considered topologically, but our method utilizes degrees of indirect correlations with seeds.

formance for language pairs with different character types.

3 Lexicon Extraction Based on Label Propagation

As described in Section 2, the performance of previous work is significantly degraded when used with a small seed lexicon. This problem could be resolved by incorporating indirect relations with all the seeds when identifying translation pairs. For example, in Figure 1, "ピラニア (piranha)" has some degree of association with the seed "魚 - fish" through "淡水 (freshwater)" in both the Japanese side and the English side, although "ピラニア (piranha)" and "魚 (fish)" do not co-occur in the same contexts. Moreover, "anaconda" has very little association with the seed "魚 - fish" in the English side. Therefore, the indirect relation with the seed "魚 - fish" helps to discriminate from between "piranha" and "anaconda" and could be an important clue for identifying a correct translation pair.

To utilize indirect relations, we introduce assumption (II): a word and its translation tend to have similar co-occurrence (direct and indirect) relations with all bilingual seeds across languages[3]. Based on assumption (II), we propose to identify a word pair as a translation pair when its co-occurrence (direct and indirect) relations with all the seeds are similar.

To obtain co-occurrence relations with all the seeds, including indirect relations, we focus on a graph-based label propagation (LP) technique (Zhu and Ghahramani, 2002). LP transfers labels from labeled data points to unlabeled data points. In the process, all vertices have soft labels that can be interpreted as label distributions. We apply LP to bilingual lexicon extraction by representing each word as a vertex in a graph with each edge encoding a direct co-occurrence relation. Translated seeds are propagated as labels, and seed distributions are obtained for each word. From the seed distributions, we identify translation pairs.

In summary, our proposed method consists of three steps (see Algorithm 1): (1) graph construction for each language, (2) seed propagation in each graph, (3) translation pair extraction.

Algorithm 1 Bilingual Lexicon Extraction

Require: comparable corpora D^e and D^f,
a seed lexicon S consists of S^e and S^f
Ensure: Output translation pairs T
1-1: $G^e = \{E^e, V^e, W^e\} \leftarrow$ construct-graph(D^e)
1-2: $G^f = \{E^f, V^f, W^f\} \leftarrow$ construct-graph(D^f)
2-1: $\tilde{G^e} = \{E^e, V^e, W^e, Q^e\} \leftarrow$ propagate-seed (G^e, S^e)
2-2: $\tilde{G^f} = \{E^f, V^f, W^f, Q^f\} \leftarrow$ propagate-seed (G^f, S^f)
3: $T \leftarrow$ extract-translation (Q^e, Q^f, S)

3.1 Graph Construction

We construct a graph representing the association between words for each language. Each graph is an undirected graph because the association does not have direction. The graphs are constructed as follows:

Step 1. Vertex assignment extracts words from each corpus, and assigns a vertex to each of the extracted words. Let $V = \{v_1, \cdots, v_n\}$ be a set of vertices.

Step 2. Edge weight calculation calculates association strength between two words as the weights of edges. Let E and W be a set of edges and that of the weights respectively, and $e_{ij} \in E$ links v_i and v_j, and $w_{ij} \in W$ is the weight of e_{ij}. Note that $|E| = |W|$.

Step 3. Edge pruning excludes edges whose weights are lower than threshold, in order to reduce the computational cost during seed propagations.

We propose two types of graphs that differ in the association measure used in Step 2: a co-occurrence graph and a similarity graph[4].

3.1.1 Co-occurrence Graph

A co-occurrence graph directly encodes assumption (II). Each edge in the graph indicates correlation strength between occurrences of two linked words. An example is shown in Figure 1.

In **edge weight calculation**, the co-occurrence frequencies are first computed for each word pair in the same context, and then the correlation strength is estimated. There are various definitions of a context or correlation measures that can be used (e.g. the

[3] Assumption (I) indicates direct co-occurrence relations between a word and its context words are preserved across different languages. Therefore, assumption (II) is derived by recursively applying assumption (I) to the "context words".

[4] We can combine the association measures used in a co-occurrence graph and a similarity graph. We will leave this combination approach for future work.

approaches used for modeling contexts in context-similarity-based methods). In this paper, we use words in a predefined window (window size is 10 in our experiments) as the context and PMI as the correlation measure:

$$w_{ij} = PMI(v_i, v_j) = log\frac{p(v_i, v_j)}{p(v_i) \cdot p(v_j)},$$

where $p(v_i)$ (or $p(v_j)$) is the probability that v_i (or v_j) occurs in a context, and $p(v_i, v_j)$ is the probability that v_i and v_j co-occur within the same context. We estimate $PMI(v_i, v_j)$ by the Bayesian method proposed by Andrade et al.(2010). Then, edges with a negative association, $PMI(v_i, v_j) \leq 0$, are pruned in **edge pruning**.

3.1.2 Similarity Graph

Co-occurrence graphs are very sensitive to accidental relation caused by lower frequent co-occurrence. Thus, we propose a similarity graph where context similarities are employed as weights of edges instead of simple co-occurrence-based correlations. Since the context similarities are computed by the global correlation among words which co-occur, a similarity graph is less subject to accidental co-occurrence. The use of a similarity graph is inspired by assumption (III): a word and its translation tend to have similar context similarities with all bilingual seeds across languages[5].

In **edge weight calculation**, we first construct a correlation vector representing co-occurrence relations for each word. The correlation vectors are constructed in the same way as the context vectors used in context-similarity-based methods (see Section 2), where context words are words in a predefined window (window size is 4 in our experiment), the association measure is PMI, and context words are treated separately for each position. A correlation vector for each position is computed separately, then concatenated into a single vector within the window. Secondly, we calculate similarities between correlation vectors. There are various similarity measures that can be used, and cosine similarity is used in this paper:

$$w_{ij} = Cos(\vec{f_i}, \vec{f_j}) = \frac{\vec{f_i} \cdot \vec{f_j}}{\|\vec{f_i}\|\|\vec{f_j}\|},$$

where $\vec{f_i}$ (or $\vec{f_j}$) is the correlation vector of v_i (or v_j). Then, in **edge pruning**, we preserve the edges with top 100 weight for each vertex.

3.2 Seed Propagation

LP is a graph-based technique which transfers the labels from labeled data to unlabeled data in order to infer labels for unlabeled data. This is primarily used when there is scarce labeled data but abundant unlabeled data. LP has been successfully applied in common natural language processing tasks such as word sense disambiguation (Niu et al., 2005; Alexandrescu and Kirchhoff, 2007), multi-class lexicon acquisition (Alexandrescu and Kirchhoff, 2007), and part-of-speech tagging (Das and Petrov, 2011). LP iteratively propagates label information from any vertex to nearby vertices through weighted edges, and then a label distribution for each vertex is generated where the weights of all labels add up to 1.

We adopt LP to obtain relations with all bilingual seeds including indirect relations by treating each seed as a label. First, each translated seed is assigned to a label, and then the labels are propagated in the graph described in Section 3.1.

The seed distribution for each word is initialized as follows:

$$q_i^0(z) = \begin{cases} 1 & if \;\; v_i \in V_s \; and \; z = v_i \\ 0 & if \;\; v_i \in V_s \; and \; z \neq v_i \\ u(z) & otherwise \end{cases},$$

where V_s is the set of vertices corresponding to translated seeds, u is a uniform distribution, $q_i^k(i = 1 \cdots |V|)$ is the seed distribution for v_i after k propagation, and $q_i^k(z)$ is the weight of a label (i.e., a translated seed) z in q_i^k.

After initialization, we iteratively propagate the seeds through weighted edges. In each propagation, seeds are probabilistically propagated from linked vertices under the condition that larger edge weights allow seeds to travel through easier. Thus, the closer vertices are, the more likely they have similar seed distributions. In Figure 1, the balloons attached to

[5]This assumption is justified because context similarities are based on co-occurrence relations that are preserved across different languages.

vertices in the graphs show examples of the seed distributions generated by propagations. For example, the English word "piranha" has the seed distribution where the weights of the seeds "Amazon", "jungle", and "fish" are 0.5, 0.3, and 0.2, respectively. Specifically, each of seed distributions is updated as follows:

$$q_i^m(z) = \begin{cases} q_i^0(z) & if \ \ v_i \in V_s \\ \dfrac{\sum_{v_j \in N(v_i)} w_{ij} \cdot q_j^{m-1}(z)}{\sum_{v_j \in N(v_i)} w_{ij}} & otherwise \end{cases},$$

where $N(v_i)$ is the set of vertices linking to v_i. We ran this procedure for 10 iterations in our experiments.

3.3 Translation Pair Extraction

After label propagations, we treat a pair of words in different languages with similar seed distributions as a translation pair. Seed distribution can be regarded as a vector where each dimension corresponds to each translated seed and each dimension has updated weight through label propagations. A similarity between seed distributions can therefore be calculated in the same way as a context-similarity-based method. In this paper, we use the cosine similarity defined by the following:

$$Cos(q_x^f, q_y^e) = \frac{\sum_{s_i \in S} q_x^f(v_i^f) \cdot q_y^e(v_i^e)}{\sqrt{\sum_{s_i \in S} (q_x^f(v_i^f))^2} \sqrt{\sum_{s_i \in S} (q_y^e(v_i^e))^2}},$$

where q_x^f (or q_y^e) is the seed distribution for a word x (or y) in the source language (or target language), S is the seed lexicon whose i-th entry s_i is a pairing of a translated seed in the source language v_i^f and one in the target language v_i^e.

4 Experiment

4.1 Experiment Data

We used English and Japanese patent documents published between 1993 and 2005 by the US Patent & Trademark Office and the Japanese Patent Office respectively, which were a part of the data used in the NTCIR-8 patent translation task (Fujii et al., 2010). Note that these documents are not aligned.

There are over three million English-Japanese parallel sentences (e.g. training data, test data, and development data used in the NTCIR-8 patent translation task, which is called *NTCIR parallel data* hereafter) in the patent data. However, a preliminary examination showed that the NTCIR parallel data covers less than 3% of all words because there are a number of technical terms and neologisms. Therefore, the patent translation task is a task that requires bilingual lexicon extraction from non-parallel data.

	Pair	Japanese Word	English Word
Lex_S	2,742	2,566	2,326
Lex_L	28,053	18,587	12,893

Table 1: Size of Seed Lexicons

We selected documents belonging to the *physics* domain from each monolingual corpus based on International Patent Classification (IPC) code[6], and then used them as a comparable corpus in our experiments. As a result, we used 1,479,831 Japanese documents and 438,227 English documents. The reason for selecting the *physics* domain is that this domain contains the most documents of all the domains.

The Japanese texts were segmented and part-of-speech tagged by ChaSen[7], and the English texts were tokenized and part-of-speech tagged by TreeTagger (Schmid, 1994). Next, function words were removed since function words with little semantic information spuriously co-occurred with many words. As a result, the number of distinct words in Japanese corpus and English corpus amounted to 1,111,302 and 4,099,825[8], respectively.

We employed seed lexicons from two sources: (1) EDR bilingual dictionary (EDR, 1990), (2) automatic word alignments generated by running GIZA++ (Och and Ney, 2003) with the NTCIR parallel data consisting of 3,190,654 parallel sentences. From each source, we extracted pairs of nouns appearing in our corpus. From (2), we excluded word pairs where the average of 2-way translation proba-

[6] *SECTION G* of IPC code indicates the *physics* domain.

[7] `http://chasen-legacy.sourceforge.jp/`

[8] The English words contain words in tables or mathematical formula but the Japanese words do not because the data format differs between English and Japanese. This is why the number of English words is larger than that of Japanese words, even though the number of English documents is smaller than that of Japanese documents.

bilities was lower than 0.5. The pairs from (1) and (2) amounted to 27,353 and 2,853 respectively, and the two sets were not exclusive. In order to measure the impact of seed lexicon size, we prepared two seed lexicons: Lex_L, a large seed lexicon that is a union of all the extracted word pairs, and Lex_S, a small seed lexicon that is a union of a random sampling one-tenth of the pairs from (1) and one-tenth of the pairs from (2). Table 1 shows the size of each seed lexicon. Note that our seed lexicons include one-to-many or many-to-one translation pairs.

We randomly selected 1,000 Japanese words as our test data which were identified as either a noun or an unknown by ChaSen and were not covered either by the EDR bilingual dictionary or by the NTCIR parallel data. This is because the purpose of our method is to complement existing bilingual dictionaries or parallel data. Note that the Japanese words in our test data may not have translation equivalents in the English side.

4.2 Competing Methods

We evaluated two types of our label propagation based methods against two baselines. $Cooc$ employs co-occurrence graphs and Sim uses similarity graphs when constructing graphs for label propagation as described in Section 3.

$Rapp$ is a typical context-similarity-based method described in Section 2 (Rapp, 1999). Context words are words in a window (window size is 10) and are treated separately for each position. Associations with context words are computed using the log-likelihood ratio (Dunning, 1993). The similarity measure between context vectors is the city-block metric.

$Andrade$ is a sophisticated method in context-similarity-based methods (Andrade et al., 2010). Context is a set of words with a positive association in a window (window size is 10). The association is calculated using the PMI estimated by a Bayesian method, and a similarity between contexts is estimated based on the number of overlapping words (see the original paper for details).

4.3 Experiment Results

Table 2 shows the performance of each method using Lex_S or Lex_L. Hereafter, $Method(L)$ (or $Method(S)$) denotes the $Method$ using Lex_L (or Lex_S). We measure the performance on bilingual lexicon extraction as Top N accuracy (Acc_N), which is the number of test words whose top N translation candidates contain a correct translation equivalent over the total number of test words (=1,000). Table 2 shows Top 1 and Top 20 accuracy. We manually[9] evaluated whether translation candidates contained a correct translation equivalent. We did not use recall because we do not know if the translation equivalents of a test word appear in the corpus.

	Lex_S		Lex_L	
	Acc_1	Acc_{20}	Acc_1	Acc_{20}
$Rapp$	1.5%	3.8%	4.8%	17.6%
$Andrade$	1.9%	4.2%	5.6%	17.6%
$Cooc$	3.2%	8.6%	9.2%	28.3%
Sim	4.1%	11.5%	10.8%	30.6%

Table 2: Performance on Bilingual Lexicon Extraction

Table 2 shows that the proposed methods outperform the baselines both when using Lex_S and using Lex_L. The improvements are statistically significant in the sign-test with 1% significance-level. The results show that capturing the relations with all the seeds including indirect relations is effective.

The accuracies of the baselines in Table 2 are worse than the previous reports: 14% Acc_1 and 46% Acc_{10} (Andrade et al., 2010), and 72% Acc_1 (Rapp, 1999). This is because previous works evaluated only the queries whose translation equivalents existed in the experiment data, which is not always true in our experiments. Moreover, previous works evaluated only high-frequency words: common nouns (Rapp, 1999) and words with a document frequency of at least 50 (Andrade et al., 2010). Our test data, on the other hand, includes many low-frequency words. It is generally true that translation of high-frequency words is much easier than that of low frequency words. We discuss the impact of test word frequencies in detail in Section 5.3.

Table 2 also shows that Sim outperforms $Cooc$ both when using Lex_S and using Lex_L. The improvements of Acc_{20} are statistically significant in the sign-test with 5% significance-level.

[9] We could not evaluate using existing dictionaries because most of the test data are technical terms and neologisms not included in the dictionaries.

	$Sim(L)$ (2)	$Cooc(L)$ (5)	$Andrade(L)$ (181)
1	psychosis	polynephropathy	disease
2	manic-depression	neuroleptic	bowel
3	epilepsy	iridocyclitis	disorder
4	insomnia	Tic	symptom
5	dementia	manic-depression	sclerosis
	$Sim(S)$ (974)	$Cooc(S)$ (1652)	$Andrade(S)$ (1747)
1	ulceration	dyslinesia	bulimia
2	ulcer	encephalomyelopathy	spasticity
3	naphthol	ganglionic	Parkinson
4	dementia	corticobasal	Asymmetric
5	gastritis	praecox	anorexia

Table 3: Translation Candidates for 躁鬱病 (manic-depression)

躁鬱病				
	$Cooc(L)$	$Andrade(L)$	$Cooc(S)$	$Andrade(S)$
1	睡眠薬 (0.12) narcotic	睡眠薬 (7.6) narcotic	痴呆 (0.016) dementia	後天 (5.0) posteriori
2	精神病 (0.11) psychosis	老年 (6.3) old	継子 (0.014) alien,stepchild	痴呆 (3.7) dementia
3	神経症 (0.08) neurosis	精神病 (6.3) psychosis	後天 (0.012) posteriori	潰瘍 (3.2) ulcer
4	ホルモン (0.05) hormone	気管支炎 (5.6) bronchitis	陽性 (0.012) electropositivity	ピリオド (2.9) period
5	不眠症 (0.04) insomnia	後天 (5.0) posteriori	潰瘍 (0.011) ulcer	重度 (2.5) seriousness
manic-depression				
	$Cooc(L)$	$Andrade(L)$	$Cooc(S)$	$Andrade(S)$
1	illness (0.15)	illness (8.6)	ganja (0.012)	galop (7.0)
2	neurosis (0.11)	psychotherapeutics (7.0)	carbanilide (0.011)	madness (5.4)
3	seizure (0.07)	galop (7.0)	paludism (0.011)	libido (5.2)
4	psychosis (0.06)	psychosis (6.8)	resignation (0.010)	vitiligo (4.6)
5	insomnia (0.04)	somnambulism (6.7)	galop (0.009)	dementia (4.3)

Table 4: Seeds with the Highest Weight

5 Discussion

5.1 Effect of Indirect Relations with Seeds

Table 3 shows a list of the top 5 translation candidates for the Japanese word "躁鬱病 (manic-depression)" for each method, where the ranks of the correct translations are shown in parentheses next to method names. Table 4 shows the top 5 translated seeds which characterize the query, where the values in parentheses indicate weight. Table 3 shows that $Cooc(L)$ can find the correct translation equivalent but $Andrade(L)$ cannot. Table 4 shows that $Cooc(L)$ can utilize more seeds closely tied to the query (e.g. "神経症 (neurosis)", "不眠症 (insomnia)"), which did not occur in the context of the query in the experiment data. The result shows that indirectly-related seeds are also important clues, and our proposed method can utilize these.

5.2 Impact of Seed Lexicon Size

Table 2 shows that a reduction of seed lexicon size degrades performance. This is natural for the baseline methods because Lex_S cannot translate most of context words, which are necessary for word characterization. Consider $Andrade(L)$ and $Andrade(S)$ in the example in Section 5.1. Table 4 shows that $Andrade(S)$ uses less relevant seeds with the query, and has to express the query by seeds with less association. For example, "精神病 (psychosis)" cannot be used in $Andrade(S)$ because Lex_S does not have the seed. Therefore, it is more difficult for $Andrade(S)$ to find correct translation pairs.

The proposed methods also share the same tendency, although each word is expressed by all the seeds in the seed lexicon. Consider $Cooc(L)$ and $Cooc(S)$ in the above example. Table 4 shows that $Cooc(S)$ expresses the query by a smooth seed distribution, which is difficult to discriminate from others. This is because Lex_S does not have relevant seeds for the query. This is why $Cooc(S)$ cannot find the correct translation equivalent. On the other hand, $Cooc(L)$ characterizes "躁鬱病" and "manic-depression" by strongly relevant seeds (e.g. "精神病 (psychosis)","神経症 (neurosis)"), and then finds the correct translation equivalent.

To examine the robustness-to-seed lexicon size, we calculated the reduction rate of Acc_{20} with the following expression: (Acc_{20} with $Lex_L - Acc_{20}$ with Lex_S) / Acc_{20} with Lex_L. The reduction rates of $Rapp$, $Andrade$, $Cooc$, and Sim are 78.4%, 76.1%, 69.6%, and 62.4% respectively. Moreover, the difference between degradation in $Cooc$ and that in $Andrade$ is statistically significant in the sign-test with 1% significance-level. These results indicate that the proposed methods are more robust to seed lexicon size than the baselines. This is because the proposed methods can utilize seeds with indirect relations while the baselines utilize only seeds in the context.

To verify our claim, we examined the number of test words which occurred with no seeds in the context. There were 570 such words in $Rapp(S)$, 387 in $Rapp(L)$, 572 in $Andrade(S)$, and 388 in $Andrade(L)$. The baselines cannot find their trans-

	Low Freq.		High Freq.	
	Acc_1	Acc_{20}	Acc_1	Acc_{20}
$Rapp(L)$	0.5%	2.4%	7.2%	25.6%
$Andrade(L)$	0.3%	1.8%	8.6%	26.3%
$Cooc(L)$	0.8%	4.3%	13.9%	40.7%
$Sim(L)$	2.2%	6.7%	15.0%	42.0%

Table 5: Comparison between Performance for High and Low Frequency Words

lation equivalents. Words such as this occur even if using Lex_L, and that number increases when Lex_S is used. On the other hand, the proposed methods are able to utilize all the seeds in order to find equivalents for words such as these. Therefore, the proposed methods work well even if the coverage of a seed lexicon is low.

5.3 Impact of Word Frequencies

Our test data includes many low-frequency words which are not covered by the EDR bilingual dictionary or the NTCIR parallel data. 624 words appear in the corpus less than 50 times. Table 5 shows Acc_N using Lex_L for 624 low-frequency words and 376 high-frequency words. Table 5 shows that performance for low-frequency words is much worse than that for high-frequency words. This is because translation of high-frequency words utilizes abundant and reliable context information, while the context information for low-frequency words is statistically unreliable. In the proposed methods, edges linking rare words are sometimes generated based on accidental co-occurrences, and then unrelated seed information is transferred through the edges. Therefore, even our label propagation based methods, especially for $Cooc$, could not identify the correct translation equivalents for rare words. Sim alleviated the problem by using a similarity graph in which edges are generated based on global correlation among words, as indicated by Table 5. Table 5 also suggests that top 20 translation candidates for high-frequency words have potential to contribute to bilingual tasks such as MT and CLIR although the overall performance is still low.

5.4 Effect of Similarity Graphs

We examined Acc_N for synonyms of translated seeds in Japanese. The Acc_1 and Acc_{20} of $Sim(L)$ are 15.6% and 56.3%, respectively, and those of $Cooc(L)$ are 9.4% and 37.5%, respectively. The results show that similarity graphs are effective for clustering synonyms into the same translation equivalents. For example, $Sim(L)$ extracted the correct translation pair of the English word "iodine" and the Japanese word "イオディン", a synonym of the translated seed "ヨウ素 (iodine)" in Japanese. This is because synonyms tend to be linked in the similarity graph and have similar seed distributions. On the other hand, in the co-occurrence graph, synonyms tend to be indirectly linked through mutual context words, so the seed distributions of the two could be far away from each other.

There are in particular many loanwords in patent documents, which are spelled in different ways from person to person. For example, the loan word for the English word "user" is often written as "ユーザ", but it is sometimes written as "ユーザー", with an additional prolonged sound mark. Therefore, Sim is particularly effective for the experiment data.

5.5 Error Analysis

We discuss errors of the proposed methods except the errors for low-frequency words (see Section 5.3). Our test data includes words whose translation equivalents inherently cannot be found. The first of these types are words whose equivalent does not exist in the English corpus. This is an unavoidable problem for methods based on comparable corpora. The second one are words whose English equivalents are compound words. The Japanese morphological analyzer tends to group a compound word into a single word, while the English text analyzer does not perform a collocation of words divided by the delimiter *space*. For example, the single Japanese word "掌紋" is equivalent to "palm pattern" or "palm print", which is composed of two words. This case was counted as an error even though the proposed methods found the word "palm" as a equivalent of "掌紋".

A main reason of errors other than those above is word sense ambiguity, which is different in every language. For example, the Japanese word "右"

means “right” and “conservatism” in English. The proposed methods merge different senses by propagating seeds through these polysemous words in only one language side. This is why translation pairs could have wrong seed distributions and then the proposed methods could not identify correct translation pairs. We will leave this word sense disambiguation problem for future work.

6 Related Work

Besides the comparable corpora approach discussed in Section 2, many alternatives have been proposed for bilingual lexicon extraction. The first is a method that finds translation pairs in parallel corpora (Wu and Xia, 1994; Fung and Church, 1994; Och and Ney, 2003). However, large parallel corpora are only available for a few language pairs and for limited domains. Moreover, even the large parallel corpora are relatively smaller than comparable corpora.

The second is a method that exploits the Web. Lu et al. (2004) extracted translation pairs by mining web anchor texts and link structures. As an alternative, mixed-language web pages are exploited by first retrieving texts including both source and target languages from the web by using a search engine or simple rules, and then extracting translation pairs from the mixed-language texts utilizing various clues: Zhang and Vines (2004) used co-occurrence statistics, Cheng et al. (2004) used co-occurrences and context similarity information, and Huang et al. (2005) used phonetic, semantic and frequency-distance features. Lin et al. (2008) proposed a method for extracting parenthetically translated terms, where a word alignment algorithm is used for establishing the correspondences between in-parenthesis and pre-parenthesis words. However, those methods cannot find translation pairs when they are not connected with each other through link structures, or when they do not co-occur in the same text.

Transliteration is a completely different way for bilingual lexicon acquisition, in which a word in one language is converted into another language using phonetic equivalence (Knight and Graehl, 1998; Karimi et al., 2011). Although machine transliteration works particularly well for proper names and loan words, it cannot be employed for phonetically dissimilar translations.

All the methods mentioned above may potentially extract translation pairs more precisely than our comparable corpora approach when their underlying assumptions are satisfied. We might improve the performance of our method by augmenting a seed lexicon with translation pairs extracted using the above methods, as experimented with in Section 4, in which additional lexical entries are included from parallel data.

7 Conclusion

We proposed a novel bilingual lexicon extraction method using label propagation for alleviating the limited seed lexicon size problem. The proposed method captures relations with all the seeds including indirect relations by propagating seed information. Moreover, we proposed using similarity graphs in propagation process in addition to co-occurrence graphs. Our experiments showed that the proposed method outperforms conventional context-similarity-based methods (Rapp, 1999; Andrade et al., 2010), and the similarity graphs improve the performance by clustering synonyms into the same translation.

We are planning to investigate the following open problems in future work: word sense disambiguation and translation of compound words as described in (Daille and Morin, 2005; Morin et al., 2007). In addition, indirect relations have also been used in other tasks, such as paraphrase acquisition from bilingual parallel corpora (Kok and Brockett, 2010). We will utilize their random walk approach or other graph-based techniques such as modified adsorption (Talukdar and Crammer, 2009) for generating seed distributions. We are also planning an end-to-end evaluation, for instance, by employing the extracted bilingual lexicon into an MT system.

Acknowledgments

We thank anonymous reviewers of EMNLP-CoNLL 2012 for helpful suggestions and comments on a first version of this paper. We also thank anonymous reviewers of First Workshop on Multilingual Modeling (MM-2012) for useful comments on this work.

References

Andrei Alexandrescu and Katrin Kirchhoff. 2007. Data-Driven Graph Construction for Semi-Supervised Graph-Based Learning in NLP. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 204–211.

Daniel Andrade, Tetsuya Nasukawa, and Junichi Tsujii. 2010. Robust Measurement and Comparison of Context Similarity for Finding Translation Pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 19–27.

Daniel Andrade, Takuya Matsuzaki, and Junichi Tsujii. 2011a. Effective Use of Dependency Structure for Bilingual Lexicon Creation. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011) - Volume Part II*, pages 80–92.

Daniel Andrade, Takuya Matsuzaki, and Junichi Tsujii. 2011b. Learning the Optimal Use of Dependency-parsing Information for Finding Translations with Comparable Corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 10–18.

Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, and Lee-Feng Chien. 2004. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1–5.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 707–718.

Dipanjan Das and Slav Petrov. 2011. Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 600–609.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT2011)*, pages 407–412.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational linguistics (COLING 2002)*, pages 1–7.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1):61–74.

EDR. 1990. Bilingual Dictionary. In *Technical Report TR-029*. Japan Electronic Dictionary Research Institute, Tokyo.

Darja Fišer, Nikola Ljubešić, Špela Vintar, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 19–26.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proceedings of the 8th NTCIR Workshop*, pages 371–376.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A New Approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, pages 1096–1102.

Pascale Fung and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 414–420.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 173–183.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL 2009)*, pages 129–137.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Herve Déjean. 2004. A Geomet-

ric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, pages 526–533.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008): the Human Language Technology Conference (HLT)*, pages 771–779.

Amir Hazem, Emmanuel Morin, and Sebastian Peña Saldarriaga. 2011. Bilingual Lexicon Extraction from Comparable Corpora as Metasearch. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 35–43.

Fei Huang, Ying Zhang, and Stephan Vogel. 2005. Mining Key Phrase Translations from Web Corpora. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, pages 483–490.

Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 481–489.

Hiroyuki Kaji. 2005. Extracting Translation Equivalents from Bilingual Comparable Corpora. *IEICE - Trans. Inf. Syst.*, E88-D:313–323.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine Transliteration Survey. *ACM Computing Surveys*, 43(3):1–46.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24:599–612.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

Stanley Kok and Chris Brockett. 2010. Hitting the Right Paraphrases in Good Time. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010)*, pages 145–153.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 617–625.

Florian Laws, Lukas Michelbacher, Beate Dorow, Christian Scheible, Ulrich Heid, and Hinrich Schütze. 2010. A Linguistically Grounded Graph Model for Bilingual Lexicon Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 614–622.

Dekang Lin, Shaojun Zhao, Benjamin Van Durme, and Marius Pasca. 2008. Mining Parenthetical Translations from the Web by Word Alignment. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008): the Human Language Technology Conference (HLT)*, pages 994–1002.

Wen-Hsiang Lu, Lee-Feng Chien, and Hsi-Jian Lee. 2004. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems*, 22(2):242–269.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 27–34.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 664–671.

Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 395–402.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Pablo Gamallo Otero and José Ramom Pichel Campos. 2008. Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2008)*, pages 423–433.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding Translations for Low-Frequency Words in Comparable Corpora. *Machine Translation*, 20:247–266.

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT2011)*, pages 1327–1335.

Reinhard Rapp. 1995. Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pages 320–322.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 519–526.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Li Shao and Hwee Tou Ng. 2004. Mining New Word Translations from Comparable Corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 618–624.

Partha Pratim Talukdar and Koby Crammer. 2009. New Regularized Algorithms for Transductive Learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2009)*, pages 442–457.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 479–484.

Dekai Wu and Xuanyin Xia. 1994. Learning an English-Chinese Lexicon from a Parallel Corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA 1994)*, pages 206–213.

Ying Zhang and Phil Vines. 2004. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. Technical report, CMU-CALD-02-107.