Decipherment with Word Embeddings through Multinomial Regression

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer
Information Sciences Institute
Department of Computer Science
University of Southern California
{qdou,knight}@isi.edu

Abstract

We introduce a base distribution informed by word embeddings into Bayesian decipherment. The learning of base distribution is combined with decipherment in a EM process. Experiments result show that the base distribution is highly beneficial to decipherment, improving deciphering accuracy from 27% to 53% for Spanish and English, 6% to 12% for Malagasy and English.

1 Introduction

Tremendous advance in Machine Translation(MT) has been made since the introduction of parallel data and machine learning techniques. However, the reliance on parallel data also slows down development and application of high quality MT systems as the amount of parallel data is often limited for low density languages and various domains.

In general, it is easier to obtain comparable monolingual data. The ability to learn translations from monolingual data could alleviate obstacles caused by insufficient parallel data. Motivated by this idea, researchers have proposed different approaches to tackle this problem.

The approaches to find translations from monolingual data can be largely divided into two groups. The first group is based on the idea proposed by Rapp (1995), where words are represented as context vectors, and two words are likely to be translations if their vectors are similar. Initially, the vectors contain just context words. Later, a number of work has extended this approach by introducing more features(Haghighi et al., 2008; Garera et al., 2009; Bergsma and Van Durme, 2011; Daumé and Jagarlamudi, 2011; Irvine and Callison-Burch, 2013b; Irvine and Callison-Burch, 2013a), using more abstract representation such as word embeddings(Klementiev et al., 2012).

Another interesting approach to solve this problem is through decipherment. It has drawn significant amount of interests in the past few years(Ravi and Knight, 2011; Nuhn et al., 2012; Dou and Knight, 2013; Ravi, 2013), and has been shown to improve machine translations. Decipherment views foreign languages as ciphers for English, and tries to find a translation table that converts foreign texts into sensible English.

Both approaches have been shown to improve quality of MT systems for domain adaptation (Daumé and Jagarlamudi, 2011; Dou and Knight, 2012; Irvine et al., 2013) and low density languages (Irvine and Callison-Burch, 2013a; Dou et al., 2014). Meanwhile, they also have their own advantages and disadvantages. While the first approach can take larger context into account, it requires high quality seed lexicons to learn a mapping between two vector spaces. In contrast, the second approach does not depend on seed lexicons, but is only able to look at limited context informed by either a bigram or trigram language model.

In this work, we take advantages of both approaches by simultaneously performing decipherment and learning a mapping between two word vector space. More specifically, we extend previous work in large scale Bayesian decipherment by introducing a better base distribution, which is informed by mapping of word embedding vectors. The main contributions of this work are:

- We propose a new framework that combines two major approaches to find translations from monolingual data.

- We show the new approach improves the state-of-the art decipherment accuracy by over two folds for multiple languages.

- We make our program a standard toolkit for finding translations from monolingual data for future research.

2 Decipherment Model

In this section, we describe the decipherment model, upon which this work is built on. We first briefly introduce recent advances made in decipherment work, then describe in details the state-of-the-art approach, and in the end, bring up the problem that we address in this work.

Unsupervised learning of translations from non-parallel is an old and challenging problem. In recent years, there has been growing interests in approaching it using decipherment techniques. Ravi and Knight (2011) built an MT system using only non parallel data for translating movie subtitles. Dou and Knight (2012) and Nuhn et al. (2012) made decipherment scalable to handle larger vocabulary. Dou and Knight (2013) improved decipherment accuracy significantly by using dependency information between words.

Throughout this paper, we use f to denote target language or ciphertext tokens, and e to denote source language or plaintext tokens. Given ciphertext $F : f_1...f_n$, the task of decipherment is to find a set of parameters $P(f_i|e_i)$ that convert F to sensible plaintext. The ciphertext F can either be full sentences (Ravi and Knight, 2011; Nuhn et al., 2012) or simply bigrams (Dou and Knight, 2013). Since using bigrams and their counts significantly speeds up decipherment, in this work, we also see F as bigrams.

Motivated by the idea from Weaver (1955), we model a ciphertext bigram F with the following generative story:

- First, a languae model $P(E)$ generates a sequence of two plaintext tokens e_1, e_2 with probability $P(e_1, e_2)$.

- Then, substitute e_1 with f_1 and e_2 with f_2 with probability $P(f_1|e_1) \cdot P(f_2|e_2)$.

Based on the above generative story, the probability of any cipher bigram F is:

$$P(F) = \sum_{e_1 e_2} P(e_1 e_2) \prod_{i=1}^{2} P(f_i|e_i)$$

Let the entire ciphertext corpus contains N such bigrams $F_1...F_N$, we write down the probability of the ciphertext corpus as:

$$P(corpus) = \prod_{j=1}^{N} P(F_j)$$

Given a plaintext bigram language model, the training objective is to find a set of parameters $P(f|e)$ that maximize $P(corpus)$. When formulated like this, one can directly apply EM to solve the problem (Knight et al., 2006). However, EM has time complexity $O(N \cdot V_e^2)$ and space complexity $O(V_f \cdot V_e)$, where V_f, V_e are the sizes of ciphertext and plaintext vocabularies respectively, and N is the number of cipher bigrams. This makes the EM approach unable to handle long ciphertext with large vocabulary size.

An alternative approach to solve the problem is to apply Bayesian inference (Ravi and Knight, 2011; Dou and Knight, 2012). Bayesian decipherment still uses the same generative story described previously. However, in Bayeisan decipherment, we no longer search for parameters $P(f|e)$ that maximize the observed ciphertext. Instead, we draw samples from plaintext sequences given the ciphertext. During sampling, the probability of any possible plaintext sample $e_1 e_2$ is computed using Equality 2:

$$P_{sample}(e_1 e_2) = P(e_1 e_2) \prod_{i=1}^{2} P_{CRP}(f_i|e_i)$$

In the above equation, the translation probability $P_{CRP}(f_i|e_i)$ is modeled by Chinese Restaurant Process(CRP), and is defined in Equation 2.

$$P_{CRP}(f_i|e_i) = \frac{\alpha P_0(f_i|e_i) + count(f_i, e_i)}{\alpha + count(e_i)}$$

where P_0 is a base distribution, also known as a prior, and α is a parameter that controls how much we trust the base distribution. $count(f_i, e_i)$ and $count(e_i)$ record the number of times f_i, e_i and e_i appear in previously generated samples respectively. The base distribution is given independently, and in all the previous work, it is set to uniform.

At the end of sampling, we compute $P(f_i|e_i)$ from ciphertext and its plaintext samples using maximum likelihood estimation:

$$P(f_i|e_i) = \frac{count(f_i, e_i)}{count(e_i)}$$

3 Model Base Distribution with Word Context Similarities

As shown in the previous section, the base distribution in Bayesian decipherment is given separately. The easiest thing to do is to set it to uniform, which is the approach taken by all previous Bayesian decipherment work. We argue that a better base distribution improves decipherment accuracy. In this work, we assign higher base distribution probabilities to word pairs that are more likely to be translations.

One straightforward way is to consider orthographic similarities. This is true for close related languages. For instance, English word “new” is translated as “neu” in German, and “nueva” in Spanish. However, this fails when two languages are not close related, such as Chinese and English.

There is a number of previous work that tries to discover translations from comparable data based on word distribution similarities. This is based on the assumption that words appear in similar context have similar meanings. The approach works very well in monolingual settings. However, when it comes to finding translations, one of the challenges is to draw a mapping between two different context space. In previous work, the mapping is usually learned from a seed lexicon.

We adopt the approach based on word context distribution similarities to learn a better base distribution. However, our work is different from previous approach in the following ways: First, our work does not rely on any seed lexicon to learn the mapping between word contexts, rather, it uses the results from sampling. Second, the mapping is not always fixed, but becomes better as the sampling process progresses. Last, but not least, the base distribution derived from the mapping and word contexts is used to improve decipherment.

4 Method

In this section, we will present our approach for learning the mapping between source and target vector spaces. Following (Mimno and McCallum, 2012), we will first derive the complete data log likelihood for our model and then present the steps of our stochastic EM algorithm. For an english word e,

	Spanish	English
Non Parallel (Gigaword)	992 million	940 million
Parallel (Europarl)	1.1 million	1.0 million

Table 1: Size of data in tokens used in Spanish-English decipherment

5 Embeddings

6 Deciphering Spanish Gigaword

In this section, we describe data and experiment details for deciphering Spanish into English.

6.1 Data

In our Spanish-English decipherment experiments, we use half of the Gigaword corpus as monolingual data, and a small amount of parallel data from Europarl for evaluation. We keep only the top 10k most frequent word types for both languages and replace all other word types with “UNK”. We also exclude sentences with more than 40 tokens as longer sentences significantly slow down the parser we use. After preprocessing, the size of data for each language is shown in Table 1. The Gigaword corpus consists of news articles from different news agencies. While we use all the monolingual data shown in Table 1 to learn word embeddings, we only parse the AFP (Agence France-Presse) section of the corpus to extract cipher dependency bigrams and build a plaintext language model. We also use GIZA(Och and Ney, 2003) to align a small amount of parallel data to build a dictionary for decipherment evaluation.

6.2 Systems

We implement a baseline system based on the work described in Dou and Knight (2013). The baseline system carries out decipherment on dependency bigrams.Therefore, we use Bohnet parser (Bohnet, 2010) to parse AFP section of both Spanish and English version of Gigaword corpus. Since not all dependency relations are shared across the two languages, we do not extract all dependency bigrams. Instead, we only use bigrams with dependency relations in the following list:

- Verb-Subject
- Verb-Noun Object

- Preposition-Preposition Object
- Noun-Noun Modifier

The baseline uses slice sampling with uniform base distribution during decipherment.

We denote the system that uses our new method as **Decipher-Embedding**. The system is the same as the baseline except that it uses a base distribution derived from word context similarities.

For all the systems, language models are built using the SRILM toolkit (Stolcke, 2002). We use modified Kneyser-Ney (Kneser and Ney, 1995) algorithm for smoothing.

6.3 Sampling Procedure

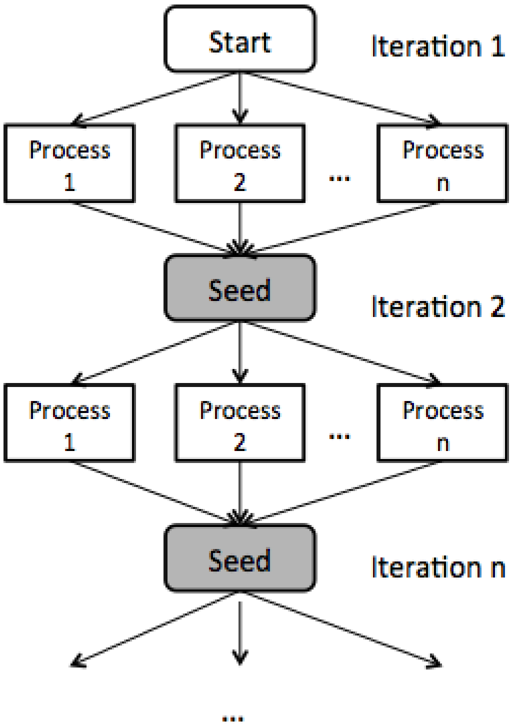

Figure 1: Iterative sampling procedures

Motivated by the previous work, we use multiple random restarts and iterative sampling process to improve decipherment (Dou and Knight, 2012). As shown in Figure. The idea is to start a few sampling processes each with a different random sample. Then combine the results from different runs and use the combined results to initiate the next sampling iteration. The details of the sampling procedure are listed below:

- Extract dependency bigrams from parsing outputs and collect their counts.
- Keep bigrams whose counts are greater than a threshold t. Then start N different randomly seeded and initialized sampling processes. Perform sampling.
- At the end of sampling, extract word translation pairs (f, e) from the final sample. Estimate translation probabilities $P(e|f)$ for each pair. Then construct a translation table by keeping translation pairs (f, e) seen in more than one decipherment and use the average $P(e|f)$ as the new translation probability.
- Lower the threshold t to include more bigrams into the sampling process. Start N different sampling processes again and initialize the first sample using the translation pairs obtained from the previous step (for each dependency bigram f_1, f_2, find an English sequence e_1, e_2, whose $P(e_1|f_1) \cdot P(e_2|f_2) \cdot P(e_1, e_2)$is the highest). Perform sampling again.
- Repeat until $t = 1$.

In our Spanish-English decipherment experiments, we use 10 different random restarts.

In experiments, we also gradually increase the weight of base distribution as more and more ciphtertext becomes available. We set the weight to 2, 4, and 6 for ciphertext with 100k, 1 million, and 10 million tokens respectively.

6.4 Evaluation Metrics

We use type accuracy as our evaluation metric: Given a word type f in Spanish, we find top K translation pairs (f, e) ranked by $P(e|f)$ from the translation table learned through decipherment. If the translation pair (f, e) can also be found in a gold translation lexicon T_{gold}, we treat the word type f as correctly deciphered. Let $|C|$ be the number of word types correctly deciphered, and $|V|$ be the total number of word types evaluated. We define type accuracy as $\frac{|C|}{|V|}$.

To create T_{gold}, we use GIZA to align a small amount of Spanish-English parallel text (1 million tokens for each language), and use the lexicon derived from the alignment as our gold translation lexicon. T_{gold} contains a subset of 4233 word types in the top 5000 frequent word types, and 7479 word types in the top 10k frequent word types. We decipher top 10k frequent Spanish word types to top 10k frequent English word types, and evaluate decipherment accuracy for both top 5k frequent word types and all the 10k word types.

7 Deciphering Malagasy

In this section, we first introduce the Malagasy language, and describe the data used in the experiments; then explain what makes deciphering Malagasy more challenging compared with Spanish, and differences in experiment settings for achieving higher decipherment accuracy.

7.1 The Malagasy Language

Malagasy is the official language of Madagascar. It has around 18 million native speakers. Although Madagascar is an African country, Malagasy belongs to the Malayo-Polynesian branch of the Austronesian language family. Malagasy and English have very different word orders. First of all, in contrast to English, which has a subject-verb-object (SVO) word order, Malagasy has a verb-object-subject (VOS) word order. Besides that, Malagasy is a typical head initial language: Determiners precede nouns, while other modifiers and relative clauses follow nouns (e.g. ny “the” boky “book” mena “red”). The significant differences in word order pose great challenges for decipherment.

7.2 Data

We list the size of both monolingual and parallel data used in this experiment in Table 2. The data used in this experiment is released from previous work by Dou et al. (2014). The monolingual data in Malagasy contains news data collected from various local websites. The English monolingual data contains Gigaword and additional 300 million tokens of news on Africa. The parallel data is collected from GlobalVoices, a multilingual news website, where volunteers translate news into different languages. The parallel data is used to build a dictionary for evaluating decipherment accuracy.

7.3 Systems

The baseline system is the same as the baseline used in Spanish-English decipherment experiments. We use data provided in previous work (Dou et al., 2014) to build a Malagasy dependency parser. For English, we use Turbo parser trained on Penn Treebank (Martins et al., 2013).

	Malagasy	English
Non Parallel	16 million (Web)	1.2 billion (Gigaword and Web)
Parallel	2.0 million (GlobalVoices)	1.8 million (GlobalVoices)

Table 2: Size of data in tokens used in Malagasy-English decipherment

Head POS	Child POS
Verb	Noun
Verb	Proper Noun
Verb	Person Pronoun
Preposition	Noun
Preposition	Proper Noun
Noun	Adjective
Noun	Determiner
Noun	Verb Particle
Noun	Verb Noun
Noun	Cardinal Number
Noun	Noun

Table 3: Head-Child POS patterns used in decipherment

Since the Malagasy parser doesn’t predict dependency relation types, we use head-child part-of-speech (POS) tag patterns to select a subset of dependency bigrams for decipherment. We list the selected POS tag patterns in Table 3.

7.4 Sampling Procedure

We use the same sampling protocol designed for Spanish-English decipherment. However, in experiments, we find out that simply using viterbi decoding to initialize the first sample does not work as well as in deciphering Spanish. Therefore, in addition to using Viterbi decoding, we also initialize the base distribution to the base distribution of previous decipherment run that produces highest decipherment accuracy.

Compared with Spanish-English decipherment, we find the base distribution plays a more important role in achieving higher decipherment accuracy for Malagasy-English. Therefore, we set weight to 10, 100, and 500 when deciphering 100k, 1 million, and 20 million ciphtertext respectively.

7.5 Results

In experiments, we gradually increase the size of ciphertext and compare decipherment accuracy of baseline with our new approach. We evaluate top 5 accuracy for top 5k and 10k most frequent word types for each language.

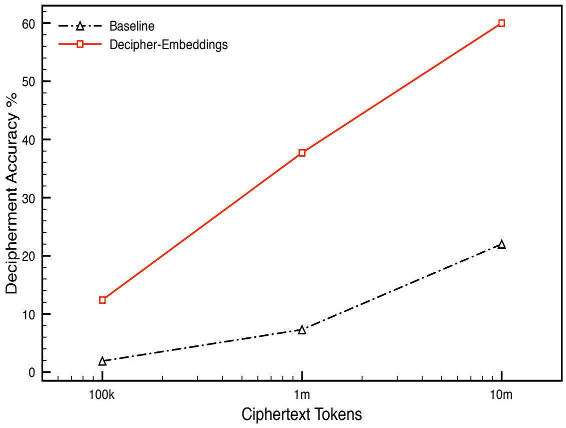

Figure 2: Learning curves for Spanish-English decipherment.

Figure 2 compares baseline with our new approach in deciphering Spanish into English. With 100k tokens of Spanish text, the baseline achieves 1.9% accuracy, while the new system achieves 12.4% accuracy, which improves the baseline by over 6 times. The improvement holds consistently throughout the experiment. In the end, the baseline achieves 22% accuracy, while the new system achieves 60% accuracy, nearly 3 folds higher.

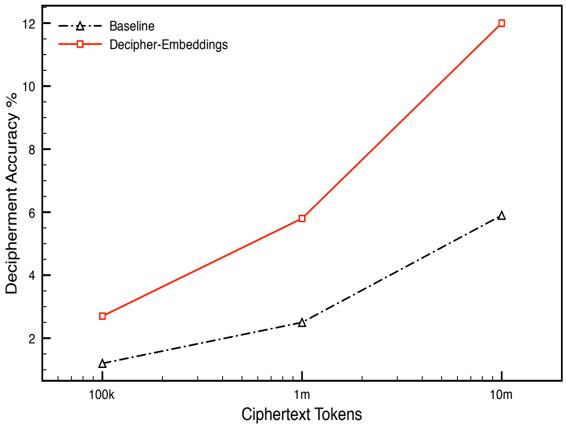

Figure 3: Learning curves for Malagasy-English decipherment.

Figure 3 compares baseline with our new approach in deciphering Malagasy into English. With 100k tokens of Malagasy data, the baseline achieves 1.2% accuracy, while the new system achieves 2.4% The improvement also holds consistently throughout the experiment. In the end, the baseline achieves 5.8% accuracy, while the new system achieves 12.0% accuracy, more than 2 times higher.

8 Conclusion and Future Work

We propose a new framework that simultaneously learns both mapping between word embeddings space and word to word translation. The approach takes advantages of two commonly used method: word vector based approach and decipherment. Experiment results show that our new algorithm improves decipherment accuracy significantly: from 22% to 60% for Spanish-English, and 5.8% to 12.0% for Malagasy-English. In the future, we will work on making the new method scale to much larger vocabulary size.

Acknowledgments

This work was supported by NSF Grant 0904684 and ARO grant W911NF-10-1-0533. The authors would like to thank David Chiang, Malte Nuhn, Victoria Fossum, Ashish Vaswani, Ulf Hermjakob, Yang Gao, and Hui Zhang (in no particular order) for their comments and suggestions.

References

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*. AAAI Press.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Coling.

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*. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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 Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013a. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, August.

Ann Irvine and Chris Callison-Burch. 2013b. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ann Irvine, Chris Quirk, and Hal Daume III. 2013. Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Sujith Ravi. 2013. Scalable decipherment for machine translation via hash sampling. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Warren Weaver, 1955. *Translation (1949). Reproduced in W.N. Locke, A.D. Booth (eds.).* MIT Press.