Measuring the Non-compositionality of Multiword Expressions

Fan Bu and **Xiaoyan Zhu**
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Sci. and Tech., Tsinghua University
buf08@mails.tsinghua.edu.cn and zxy-dcs@tsinghua.edu.cn

Ming Li
David R. Cheriton School of Computer Science
University of Waterloo
mli@uwaterloo.ca

Abstract

Multiword Expressions (MWEs) appear frequently and ungrammatically in the natural languages. Identifying MWEs in free texts is a very challenging problem.

This paper proposes a knowledge-free, training-free, and language-independent Multiword Expression Distance (MED). The new metric is derived from an accepted physical principle, measures the distance from an n-gram to its semantics, and outperforms other state-of-the-art methods on MWEs in two applications: question answering and named entity extraction.

1 Introduction

A Multiword Expression (MWE) is a sequence of neighboring words "whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components" (Choueka, 1988). In the paper, MWEs refer to non-compositional lexical units including idioms, terminologies and name entities. As Jackendoff (1997) notes, the magnitude of MWEs is far greater than what has traditionally been realized within linguistics. He estimates that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. In WordNet 1.7 (Fellbaum, 1998), 41 percent of the entries are multi-words. Some specialized domain vocabulary, such as terminology, overwhelmingly consists of MWEs. Automatic extraction of MWEs is indispensable to many tasks such as machine translation, name entity extraction, information retrieval and question answering.

Due to their non-compositionality, many MWEs cannot be directly identified using grammatical rules, which poses a major challenge to automatic analysis. Moreover, existing resources like dictionaries can never have adequate and timely coverage. Therefore people turn to statistical method to characterize MWEs.

Since Church and Hanks (1990) proposed Pointwise Mutual Information (PMI), a variety of measures, such as Log-likelihood, Symmetrical Conditional Probability (SCP) and Mutual Expectation (Dias et al., 2000), have been introduced to measure word association. Their basic ideas are very similar: the whole n-gram is separated into two parts and the association is determined by the joint probability and the probability of each part. Pecina (2006) compared 84 bi-gram association measures and found PMI has the best performance in Czech data. When applying these measures to the n-grams for $n > 2$, it is not clear how can the association between the deliberately separated two parts represent the non-compositionality of the whole n-gram. Different policies have been studied to extend these measures into arbitrary n-grams (Silva and Lopes, 1999; Schone and Jurafsky, 2001; Dias et al., 2000). Is there a fundamental, less arbitrary, and general approach to this problem? That is,

- Can we actually derive a MWE metric for n-grams from the first principles, instead of making a seemingly sensible, but really arbitrary, proposal?
- Will such a theoretically justified new metric actually works better than other heuristic

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 116–124,
Beijing, August 2010

measures for general MWEs?

This paper will answer above questions positively. We derive an optimal distance metric Multiword Expression Distance (MED). MED defines the semantic function for n-grams and the information distance (Bennett et al., 1998) from the n-grams to their semantics. Unlike previous methods it ensures the cohesion of the n-gram directly hence applicable to MWEs of any length.

The MED is naturally generalized to its conditional version. The extension is based on the observation that many MWEs are domain dependent. It is true that some MWEs are only used in certain domains, but they are domain free. For example, we know that "polymerase chain reaction" is some sort of terminology even if many of us do not know what it is exactly. However that is not always the case. For those who do not watch movies, the sentence "catch me if you can" will probably be taken as a non-MWE, instead of a movie name. The non-compositionality of this sentence appears only in the movies domain. The experimental results show that given appropriate phrases as conditions, the conditional MED performs better than MED.

We also investigate the efficacy of MED on post-processing of Question Answering (QA) and complex named entity extraction. The experimental results show that our method outperforms state of art methods (Zhang et al., 2009; Downey et al., 2007) in these two applications. Moreover, MED is a pure statistical metric which can be easily combined with other methods.

The remainder of this paper is organized as follows: In the next section we review the related work on Multiword Expression and information distance. Section 3 gives a preliminary introduction to Kolomogorov complexity and information distance. Section 4 proposes the formal definition of MED. In Section 5 we discuss the difference between MED and Pointwise Mutual Information. We apply MED to QA post-processing and complex named entity extraction in Section 6 and evaluate their performance in Section 7. In the last section we conclude this work.

2 Related Work

Researchers have explored various techniques for identifying MWEs. These approaches could be broadly classified into three types: linguistic methods, sequential tagging based methods and statistical methods.

The mostly used linguistic information for MWE extraction is words' Part-Of-Speech tags. Justeson and Katz (1995) extracted technical terminologies from documents using a regular expression on POS-tags of a word sequence, together with some frequency constraints. Argamon et al. (1998) separated the POS sequence of a multi-word into small POS tiles, counted tile frequency in the MWE and non-MWE training sets and identify new MWEs by these counts. Although linguistic methods perform well in term extraction on specific domains, it cannot be generalized to identify arbitrary MWEs.

Several supervised learning methods have been used previously for extracting Name Entities including Hidden Markov Models, Maximum Entropy Markov Models and Conditional Random Field (CRF) models (McCallum and Li, 2003). In order to allow tractable computation, these models can only use local features in a small window. Although the approximate inference methods have been incorporated into sequential tagging model to capture non-local information (Finkel et al., 2005), these models are not capable of recognizing complex named entities, especially those containing conjunctions and prepositions. Experimental results in (Downey et al., 2007) show that statistical methods substantially outperform sequential tagging based methods on identifying complex named entities.

In statistical methods for MWE extraction, Church and Hanks (1990) first presented Pointwise Mutual Information (PMI) as an objective measure for estimating word association. Since then, many methods has been proposed to measure bi-gram association, such as Log-likelihood (Dunning, 1993) and Symmetrical Conditional Probability (Silva and Lopes, 1999). Pecina (2006) compared 84 bi-gram association measures and concluded that PMI had the best performance in Czech data. When it comes to measure

the non-compositionality for arbitrary n-grams, policies were taken to separate n-gram into two parts X and Y so that it can be measured by existing bi-gram methods (such as PMI). Silva and Lopes (1999) and Dias et al. (2000) calculated the arithmetic average of every possible separation. Schone and Jurafsky (2001) define X and Y to be the word sequences $w_1 w_1 ... w_i$ and $w_{i+1} w_{i+2} ... w_n$, where i is chosen to maximize $P_x P_y$. Recently Zhang et al. (2009) proposed Enhanced Mutual Information (EMI) which measured the cohesion of n-gram by the frequency of itself and the frequency of each word.

The information distance is a universal distance measure between two information carrying entities (Bennett et al., 1998; Li et al., 2001; Li et al., 2004). The applications of information distance using compression were first introduced in (Li et al., 2001) and then in (Bennett et al., 2003; Chen et al., 2004). The experimental results in (Keogh et al., 2004) showed that information distance/compression based method was superior to 51 parameter-laden methods from seven major data mining conferences on their benchmark data. The web-based approximation of information distance was introduced by Cilibrasi and Vitányi (2007) to measure the semantic similarity of two words or concepts.

3 Preliminaries

3.1 Kolmogorov Complexity

Kolmogorov complexity defines randomness of an individual string. Fix a universal Turing machine U, the *Kolmogorov complexity* of a binary string x condition to another binary string y $K_U(x|y)$ is defined as the length of the shortest (prefix-free) program for U that outputs x with input y. It can be shown that for a different universal Turing machine U', for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C, \qquad (1)$$

where the constant C depends only on U'. Thus, we can simply write $K_U(x|y)$ as $K(x|y)$ and $K(x|\epsilon)$ as $K(x)$, where ϵ is the empty string.

3.2 Information Distance

Between any two information carrying entities, is there an objective distance that is application-independent and unique, similar to the concept of distance in the physical world? From a commonly accepted physical principle of von Neumann and Landauer that irreversibly processing one bit of information costs 1KT of energy, Bennett et al. (1998) derived exactly such a distance: the Information Distance. Information Distance $E(x, y)$ between two objects x and y is the energy to convert between x and y. Bennett et al. (1998) proved:

Theorem 1 *Up to an additive logarithmic term,* $E(x, y) = \max\{K(x|y), K(y|x)\}$.

Thus, the max distance was defined below (Bennett et al., 1998):

$$D_{max}(x, y) = \max\{K(x|y), K(y|x)\}.$$

D_{max} was shown to satisfy distance requirements such as positivity, symmetricity and triangle inequality (Bennett et al., 1998). It was further shown that D_{max} is optimal in the sense that it is universal. That is, it minorizes (up to constant factors) all other nontrivial and computable distances. More precisely, a distance *D* is admissible if

$$\sum_y 2^{-D(x,y)} \leq 1. \qquad (2)$$

Thus, we exclude trivial distances such as $d(x, y) = 0$ for all x, y. It was proved in (Bennett et al., 1998) that for any admissible computable distance *D*, there is a constant *c*, for all *x*,*y*,

$$D_{max}(x, y) \leq D(x, y) + c.$$

In other words, if any such distance D discovers some similarity between x and y, so will D_{max}.

In order to deal with the information carrying objects of different sizes, the normalized information distance was proposed in (Li et al., 2001). In (Li et al., 2004), the normalized max distance was defined as:

$$d_{max}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$

d_{max} satisfies positivity, symmetricity, triangle inequality and some weak form of universality (Li et al., 2004).

4 A New Metric for MWE

4.1 The Semantics

When applying the Information Distance to identifying MWEs, how to encode n-grams and their semantics is the first thing to be considered. It is inappropriate to encode MWEs literally. For example, when referring to "kick the bucket", the three words "kick", "the" and "bucket" cannot represent all the semantics about this expression.

Inspired by Cilibrasi and Vitányi (2007), we define *context* of an n-gram as the set of all the web pages containing it. Also, *semantic* of an n-gram is defined as the set of all the web pages containing all the words appeared in that n-gram. For example, the semantic of "U.S. president" including not only the pages containing itself but also those containing "the president of U.S." or "president Obama says that ... U.S. government...".

4.2 Multiword Expression Distance

Let us denote the vocabulary set by S and the set of web pages by Ω. The cardinality of Ω is denoted by M=$|\Omega|$. Define $G \equiv S^+$ as the set of n-grams. A search term t is defined as an n-gram or the conjunction of search terms. Denote T as the set of search terms and we have $G \subset T$. Let $\phi : T \to 2^\Omega$ be the *context* function mapping each search term t to the web set which includes (and only includes) all the web pages containing all the n-grams in t. Let $\theta : G \to T$ be the function mapping each n-gram $g = w_1 w_2 ... w_n$ to $\bigwedge_i w_i$, the conjunction of the words in it. Finally we define the *semantic* function $\mu : G \to 2^\Omega$ as the composite function $\phi \circ \theta$. It is obvious that for any n-gram g, we have $\phi(g) \subseteq \mu(g)$. Given an n-gram g, we will encode $\phi(g)$ and $\mu(g)$ and calculate the distance between them.

While $K(x)$ is not computable, a simple heuristic, noticed by Cilibrasi and Vitányi (2007), is to use Shannon-Fano code to encode the probability (approximated by its internet frequency) of x. Assume that all web pages are equiprobable, with the probability of being returned by search engine being $\frac{1}{M}$. Let $p : \phi(T) \to [0, 1]$ be the *context* probability function in which $\phi(T) \equiv \{x | \exists y \in T, x = \phi(y)\}$. Since each context is a set of webpages, the probability of context c is defined as $p(c) = \frac{|c|}{N}$ where $N = \sum_{c \in \phi(T)} |c|$ ensures p is a valid probability function. The Shannon-Fano code (Li and Vitányi, 2008) length associated with p can then be regarded as an approximation of K,

$$K(x) \approx -\log p(x) \tag{3}$$

$$K(x, y) \approx -\log p(x, y) \tag{4}$$

According to (3),(4) and Theorem 1, D_{max} can be approximated as follows:

$$\begin{aligned} & D_{max}(x, y) \\ = \; & \max\{K(x|y), K(y|x)\} \\ = \; & K(x, y) - \min\{K(y), K(x)\} \\ \approx \; & \max\{\log|x|, \log|y|\} - \log|x \cap y| \end{aligned}$$

Similarly, we have

$$\begin{aligned} & D_{max}(x, y|c) \\ \approx \; & \max\{\log|x \bigcap c|, \log|y \bigcap c|\} - \log|x \bigcap y \bigcap c| \end{aligned}$$

Since $\phi(g) \subseteq \mu(g)$, the Multiword Expression Distance of an n-gram g can be defined as follows:

$$\begin{aligned} & \mathrm{MED}(g) \\ \equiv \; & D_{max}(\phi(g), \mu(g)) \\ \approx \; & \max\{\log \tfrac{|\phi(g)|}{|\phi(g) \bigcap \mu(g)|}, \log \tfrac{|\mu(g)|}{|\phi(g) \bigcap \mu(g)|}\} \\ = \; & \log|\mu(g)| - \log|\phi(g)| \end{aligned}$$

Given a search term c as condition, the Conditional Multiword Expression Distance of an n-gram g is defined as follows:

$$\begin{aligned} \mathrm{MED}(g|c) \; \equiv \; & D_{max}(\phi(g), \mu(g)|\phi(c)) \\ \approx \; & \log|\mu(g) \bigcap \phi(c)| - \log|\phi(g) \bigcap \phi(c)| \end{aligned}$$

Based normalized information distance, NMED and its conditional version can be derived as follows:

$$\begin{aligned} \mathrm{NMED}(g) \; & \approx \; \tfrac{\log|\mu(g)| - \log|\phi(g)|}{\log N - \log|\phi(g)|} \\ \mathrm{NMED}(g|c) \; & \approx \; \tfrac{\log|\mu(g) \bigcap \phi(c)| - \log|\phi(g) \bigcap \phi(c)|}{\log|\phi(c)| - \log|\phi(g) \bigcap \phi(c)|} \end{aligned}$$

Where N can be estimated from the size of internet by some combinatorial methods.

To implement MED by a general search engine, we assume Ω to be the set of indexed webpages. Thus, $|\phi(g)|$ and $|\mu(g)|$ can be approximated by the hit numbers given g and the "logic and" of each word in g as queries. Yahoo Search is used in our experiments.

5 Relation with Pointwise Mutual Information

When $n = 2$, we denote $\mathrm{P}(w_1w_2)$ the probability of a web page containing bi-gram $g = w_1w_2$ and $\mathrm{P}(w_1 \bigwedge w_2)$ the probability of a web page containing w_1 and w_2. Assuming the occurrence of w_1 and w_2 are independent, we have

$$\begin{aligned} \mathrm{MED}_2(g) &= \log \frac{|\phi(w_1 \bigwedge w_2)|}{|\phi(w_1w_2)|} \\ &= \log \frac{\mathrm{P}(w_1 \bigwedge w_2)}{\mathrm{P}(w_1w_2)} \\ &\approx \log \frac{\mathrm{P}(w_1)P(w_2)}{\mathrm{P}(w_1w_2)} \\ &\propto -\mathrm{PMI}(g) \end{aligned}$$

Thus, PMI is inversely proportional to MED under the independence assumption. This assumption is unadvisable for obvious reasons. PMI compares the probability of observing x and y within a given window w (w=2 when measuring collocation) with the probabilities of observing x and y independently. However, most of the word sequences in practice (both MWEs and non-MWEs) are far from being independent. Therefore the assumption potentially creates additional noises to MED, especially when $n > 2$. The internet contains billions of pages and thus we can count the pages containing specified words directly without making independent assumption to overcome data sparseness.

6 Applications

6.1 MWE for QA Systems

Some types of questions require a QA system to return phrases as the answers instead of sentences, such as Factoid and List. Given a question, we need to generate queries, obtain relevant pages from the internet, extract the candidate n-grams from relevant pages and finally rank all the candidates by their likelihood of being an answer.

Some previous work exploited web redundancy to estimate answer validity(Magnini et al., 2002; Zhang et al., 2008). No research, to our knowledge, has focused on checking the completeness of candidates. Most of texts on the internet are informal (e.g. they contain uncapitalized proper nouns and incomplete sentence structures). Parser and named entity recognizers trained on formal corpus are unpractical on recognize NP chunks or name entities on the web.

Observing that each candidate is n-gram and checking the completeness of a candidate is to measure its non-compositionality, we introduce a simple MWEs-based method to rank all candidates by their completeness and merge similar answers.

Given a question and a list of candidate answers:

1. Extract proper nouns from the question as conditions.
2. Calculate the conditional MED (or MED if no proper noun is found in question) for each candidate. Then for each pair of literally similar candidates, the one with larger MED distance is removed.
3. Rank the rest candidates by conditional MED.

This method is case insensitive and do not rely on context information. All of the statistics are performed on the internet thus no local corpus is needed.

6.2 Complex Named Entity Extraction

In many previous work (McCallum and Li, 2003; Finkel et al., 2005), named entity extraction is combined with classification, which is known as Name Entity Recognition (NER). Most of these NER technique are based on sequential tagging models and unsuitable to the task of locating complex named entities in Web text. In (Downey et al., 2007), the author treated named entity as a type of MWE and proposed the algorithm LEX++ to locate complex named entities.

Inspired by Downey's work, we propose a conditional MED based algorithm MWE++ to extract named entities. Given a sentence $S = \{S_1, S_2...S_n\}$ and parameters τ_1,τ_2 and δ, MWE++ proceeds as follows:

1. Initialize a sequence of names $N = (n_1, n_2, ..., n_M)$ equal to the maximal contiguous substrings of S that consist entirely of capitalized words. If the first word of S appears capitalized in the local corpus and it is at the beginning of a sentence more than δ of the times, it is omitted from N.

2. Until N does not change during last iteration:

   (a) Choose the *mergeable* pair of names (n_i, n_{i+1}) with minimum conditional MED.

   (b) Replace n_{min_i} and n_{min_i+1} with the single name $n_{min_i} w_{min_i} n_{min_i+1}$ where w_i is the uncapitalized words between n_i and n_{i+1}.

3. For every names n_i in N

   (a) Check common prefix and punctuation at boundary of n_i via local corpus.

   (b) Check number at boundary of n_i via internet.

In MWE++, We define two thresholds τ_1 and τ_2 to estimate the name entity confidence of a given n-gram. If $\mathrm{MED}(g|.)$ is lower than τ_1, between τ_1 and τ_2 or higher than τ_2, $\mathrm{conf}(g)$ will be 2 (Definitely), 1 (Probably) or 0 (Impossible). The confidence of all initialized capitalized words will be set to 1. If an n-gram contain unmatched brackets or quotation marks, its confidence will be set to 0. Also, The confidence of n-gram containing comma will be reduced by 1. We say a pair of names (n_i, n_{i+1}) is mergeable if and only if $\mathrm{conf}(n_i w_i n_{i+1}) \geq \max(\mathrm{conf}(n_i), \mathrm{conf}(n_{i+1}))$.

After iteration, we will check common prefixes, punctuations and numbers at boundary of each names. If a name n_i is immediately preceded by a single number t and $\mathrm{conf}(tn_i) \geq 1$, we replace n_i by tn_i. Similarly, a number t immediately following n_i is appended to n_i when $\mathrm{conf}(n_i t) \geq 1$. Due to the limitation of search engine, punctuation check and common prefix check modules are performed on local corpus just the same as LEX++.

7 Experiments and Analysis

7.1 Compositionality Measure

In this section, we evaluate how well can MED separate non-compositional phrases (idioms) from compositional ones. First we evaluate MED and other four metrics on English_VPC data published on the MWE 2008 shared task. The data set contains 3078 verb-noun bi-grams and 14 percent of them are annotated as idiomatic. The average precision of MED, PMI, SCP, t-score and EMI (Zhang et al., 2009) are 0.234, 0.233, 0.285, 0.274 and 0.205. The result shows that MED is not distinguished on bi-grams test. It is partly because most idiomatic verb-noun collocations are often used non-idiomatically. Their compositionality are not necessarily lower than non-idiomatic ones.

We also evaluate different metrics on n-grams of varied lengths. Since all published MWE data sets we find only contain bi-grams, we construct our test set as follows. We first collected common idioms from the lists of english idioms on Wikipedia. To get enough common but not idiomatic phrases, we collect common compositional phrases from UsingEnglish.com, englishspeak.com, Wikipedia and China Daily BBS. Since it is difficult for non-native speakers to pick up idioms from non-idiomatic ones, we do not manually check all compositional phrases. The test set contains 1529 idioms and 1798 compositional phrases. The n-gram frequencies are not significantly different between idioms and compositional phrases. The mean and standard deviation are 2.1×10^5 and 7.8×10^5 on idioms and 7.4×10^5 and 4.8×10^6 on compositional phrases. We employ different measures to rank all the phrases. Non-conditional MED and NMED are compared with AVG_SCP (Silva and Lopes, 1999), MAX_PMI (Schone and Jurafsky, 2001), EMI (Zhang et al., 2009) and the baseline n-gram frequency. T-score is not under evaluation because we do not find sound n-gram extension for it. The precision-recall curve is shown in Fig. 1. Since the performance of MED and NMED are very

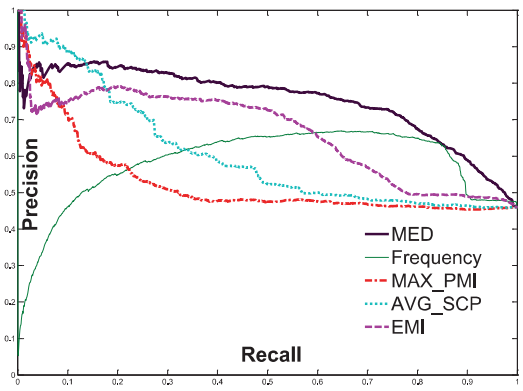

Figure 1: Precision-recall curves of five measures

	freq	MAX_PMI	AVG_SCP	EMI	NMED	MED	MED(.|.)
fairy tale	0.493	0.484	0.570	0.515	0.615	0.617	**0.657**
science fiction	0.500	0.470	0.558	0.525	0.596	0.599	**0.633**
action movie	0.695	0.523	0.723	0.703	0.763	0.768	**0.823**
animation	0.561	0.642	**0.693**	0.489	0.671	0.673	0.689
horror movie	0.595	0.528	0.647	0.633	0.667	0.670	**0.692**
documentary	0.525	0.549	0.626	0.512	0.596	0.598	**0.654**
hip hop	0.598	0.627	0.645	0.635	0.652	0.651	**0.712**
jazz	0.549	0.501	0.543	0.539	0.627	0.625	**0.716**
rock&roll	0.742	0.567	0.730	0.741	0.708	0.717	**0.836**
company	0.614	0.584	0.689	0.663	0.754	**0.756**	0.735
soccer player	0.945	0.648	0.904	**0.973**	0.911	0.918	0.941
novelists	0.772	0.701	**0.870**	0.866	0.821	0.828	0.864
PS3 game	0.603	0.675	0.740	0.535	0.742	**0.744**	0.727
overall	0.612	0.577	0.688	0.629	0.696	0.700	**0.726**

Table 1: Performance of different measures in each list

close, NMED is not displayed for clarity. From the result we can see that MED performs substantially better than all the other measures. Average precision(avp) of the top 3 measures MED, EMI and AVG_SCP are 0.75, 0.71 and 0.66.

7.2 QA Post-processing

It is difficult to evaluate the method introduced in Section 6.1 directly since QA benchmarks mainly focus on accuracy of the top one answer instead of the completeness of top-n candidates. Therefore, the experiment is designed as follows. We extract name lists on different domains from Wikipedia. For each name in each list, we put it into a search engine and get the context from a random selected snippet. For each name, We created two incomplete names by randomly adding (or removing) one or two words according to its context. It is guaranteed that the original name and its counterpart with noise must have at least two words in common. We tag the original names and the noise added ones in each list as positive and negative samples. A list can be regarded as the candidates and the list name (or its synonym) can be seen as the key phrase extracted from question.

The test set can be divided into six common categories: movie, book, music, person, organization and video game. Each category contains one to four lists. The test set contains 11080 samples in total. Still, we employ the measures in previous experiments to rank all the candidates to see if the complete names can be separated from the incomplete names. The results are listed in Table 1. The overall avp is the average of the avp of each lists weighted by their size.

It is shown that the performance of conditional MED is the best over all metrics, followed by MED. The reason why EMI and AVG_SCP get best results on soccer player and novelists is that they take more advantage of frequency. Since the length of people's name are short (2 to 3 words), most of negative samples are created by adding words, which makes frequency important.

7.3 Complex Named Entity Extraction

In this section we evaluate the named entity extraction performance of Algorithm MWE++. The experiment is done on the corpus, the training set and the test set provided by Downey et al. (2007). Four classes of entities (Actor, Book, Company and Film) were manually annotated on both training and test set. All sentences in the corpus contain named entities from the above four classes (but not annotated). The corpus consists of 183,726 sentences while the training and the test set contain 200 and 629 sentences, respectively. Furthermore, test sentences are separated into 100 difficult cases and 529 easy cases. All difficult cases contain complex name entities (entities containing uncapitalized words), such as "Procter and

Gamble" and "Gone with the Wind".

The conditional MED metric in this experiment is redefined as follows:

$$\mathrm{MED}(g|C) = \min_{c \in C}\{\mathrm{MED}(g|c)\},$$

where C={"IMDB","Amazon","corporation"}. "IMDB" is used as the condition of Actor and Film while "Amazon" and "corporation" are chosen to be the condition of Book and Company. We compute the conditional MED for all entities on training set. τ_1 is set to the median and τ_2 is set to the value larger than 90% entities on training set. δ is set to 0.5. MWE++ is performed on the 100 difficult cases. The results shown in Table 2 convincingly show that MWE++ significantly outperforms LEX++, supervised models (SVMCMM, CRF) and rule-based model (MAN) on identifying complex named entities. Compared to LEX++, MWE++ is not only more accurate but also more flexible. LEX++ relies on local corpus while MWE++ does not. When recognizing new entities, we just need to find appropriate condition words instead of preparing new corpus. For the sake of completeness, the F-score of MWE++ on easy cases is 91, which is lower than all the other methods. However this is irrelevant since this part can be made quite accurate by specialized databases and training by any known methods.

All test data in this paper can be downloaded from http://60.195.250.61:8080/download/.

8 Conclusion

We have derived an MWE metric MED from the first principles via Information Distance. The new metric measures the distance from an n-gram to its semantics. It is provably optimal (universal), overcomes several deficiencies of previous approaches, and convincingly outperforms the other methods.

Also, we have taken advantage of the fact that some MWEs are domain dependent. This feature is important when recognizing named entities and terminologies. The conditional MED is better than MED when we know what we are looking for. Since MED is quite different from previous measures, it can be combined with others by machine learning approaches and enhance the overall performance. Further experiments are needed.

	F_1	Recall	Precision
MAN	0.18	0.22	0.16
CRF	0.35	0.42	0.31
SVMCMM	0.42	0.48	0.37
LEX++	0.74	0.76	0.72
MWE++	**0.83**	**0.86**	**0.80**

Table 2: Named entity extraction on difficult cases

Acknowledgment

This work was supported mainly by Canada's IDRC Research Chair in Information Technology program, Project Number: 104519-006. It is also supported by the Chinese Natural Science Foundation grant No. 60973104, NSERC Grant OGP0046506, 863 Grant 2008AA02Z313 from China's Ministry of Science and Technology, Canada Research Chair program, MITACS, an NSERC Collaborative Grant, and Ontario's Premier's Discovery Award.

References

Shlomo Argamon, Ido Dagan and Yuval Krymolowski 1998. A memory-based approach to learning shallow natural language patterns. In *Proc. of COLING*, 1998, pp. 67-73.

Charles H. Bennett, Peter Gács, Ming Li, Paul M.B. Vitányi, and Wojciech H. Zurek 1998. Information distance. *IEEE Trans-IT* 44:4, 1998, pp. 1407-1423.

Charles H. Bennett, Ming Li and Bin Ma 2003. Chain letters and evolutionary histories. *Scientific American*, 288:6, (feature article), 76-81.

Xin Chen, Brent Francia, Ming Li, Brian Mckinnon, Amit Seker 2004. Shared information and program plagiarism detection. *IEEE Trans. Information Theory*, 50:7, 1545-1550.

Yaacov Choueka 1988. Looking for needles in a haystack or locating interesting collocation expressions in large textual databases In *Proc. of the RIAO*, 1988, pp. 38-43.

Kenneth W. Church and Patrick Hanks 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.

Rudi L. Cilibrasi and Paul M.B. Vitányi 2007. The Google similarity distance. *IEEE Trans-Knowledge and Data Engineering* 19:3, 2007, pp. 370-383.

Joaquim F. da Silva and Gabriel P. Lopes 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proc. of Sixth Meeting on Mathematics of Language*, pp. 369-381.

Gaël Dias, Sylvie Guilloré and José G.P. Lopes 2000. Mining textual associations in text corpora. In *Sixth ACM SIGKDD, Workshop on Text Mining*, pp. 92-95.

Ted Dunning 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. Vol. 19, No. 1, 61-74.

Doug Downey, Matthew Broadhead and Oren Etzioni 2007. Locating complex named entities in web text. In *Proc. of IJCAI*, 2007, pp. 2733-2739.

Christine Fellbaum 1998. WordNet: an electronic lexical database. MIT Press, Cambrige, MA.

Jenny R. Finkel, Trond Grenager and Christopher Manning 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of ACL*, 2005, pp. 363-370.

Ray Jackendoff 1997. The architecture of the language faculty. MIT Press, Cambrige, MA.

John S. Justeson and Slava M. Katz 1995. Technical terminology: Some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering* 1:1, 1995, 9-27.

Eamonn J. Keogh, Stefano Lonardi and Chotirat A. Ratanamahatana 2004. Towards parameter-free data mining. In *Proc. of ACM SIGKDD*,2004, pp. 206-215.

Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:2(2001), 149-154.

Ming Li, Xin Chen, Xin Li, Bin Ma and Paul M.B. Vitányi 2004. The similarity metric. *IEEE Trans-IT* 50:12, 2004, 3250-3264.

Ming Li and Paul M.B. Vitányi 2008. An introduction to kolmogorov complexity and its applications. Springer-Verlag, New York, 2008. Third edition.

Andrew McCallum and Wei Li 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of the 7th conference on Natural language learning at HLT-NAACL*,2003 ,pp. 188-191.

Bernardo Magnini, Matteo Negri and Hristo Tanev 2002. Is it the right answer? Exploiting web redundancy for answer validation. In *Proc. of ACL*,2002, pp. 425-432.

Pavel Pecina 2006. An extensive empirical study of collocation extraction methods. In *Proc. of COLING-ACL*, 2006, pp. 953-960.

Patrick Schone and Daniel Jurafsky 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proc. of EMNLP, 2001.*

Wen Zhang, Taketoshi Yoshida, Xijin Tang and Tu-Bao Ho 2009. Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Systems with Applications*, Volume 36, 10919-10930, Elsevier.

Xian Zhang, Yu Hao, Xiaoyan Zhu and Ming Li 2007. New information measure and its application in question answering system. *J. Comput. Sci. Tech.*, 23:4(2008), pp. 557-572. (Preliminary version appeared in SIGKDD 2007.)