Chapter 10

TRANSLINGUAL MINING FROM TEXT DATA

Jian-Yun Nie
University of Montreal
Montreal, H3C 3J7, Quebec, Canada
nie@iro.umontreal.ca

Jianfeng Gao
Microsoft Corporation
Redmond, WA, USA
jfgao@microsoft.com

Guihong Cao
Microsoft Corporation
Redmond, WA, USA
gucao@microsoft.com

Abstract Like full-text translation, cross-language information retrieval (CLIR) is a task that requires some form of knowledge transfer across languages. Although robust translation resources are critical for constructing high quality translation tools, manually constructed resources are limited both in their coverage and in their adaptability to a wide range of applications. Automatic mining of translingual knowledge makes it possible to complement hand-curated resources. This chapter describes a growing body of work that seeks to mine translingual knowledge from text data, in particular, data found on the Web. We review a number of mining and filtering strategies, and consider them in the context of statistical machine translation, showing that these techniques can be effective in collecting large quantities of translingual knowledge necessary for CLIR.

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_10,

Keywords: cross-lingual mining, translingual mining, cross-lingual information retrieval

1. Introduction

The principle goal of text mining is to discover knowledge from text data. Various forms of knowledge may be involved, including possibly concepts and relations among them. While the bulk of work on text mining has been conducted on monolingual texts, relating to identifying concepts and relations among them in a single language, a by-no-means negligible class of applications involves more than one language. The prototypical member of this class is Machine translation (MT), which seeks to transfer a sentence or a text from a language into another. To do this, one has to create or extract various types of *translingual* knowledge such as word translation (usually in the form of a bilingual dictionary or a statistical translation model) and methods of syntactic transfer. Whereas classical MT systems were once constructed using manually defined rules and dictionaries, modern MT systems exploit large bilingual text data from which to obtain translational knowledge automatically. The extraction of this translational knowledge is, in its essence, a form of translingual text mining. Another important application that calls for translingual text mining is cross-language information retrieval (CLIR), in which one tries to retrieve documents in a language different from the language of the original query. A person may wish, for example, to retrieve documents in English using a query in Chinese. Although additional translational knowledge may need to be brought to bear in order to compare the returned documents and the query in two languages, the informational goal of CLIR is distinct from that of full text MT, and the process of extracting translingual knowledge differs accordingly.

In this chapter, we survey some of the approaches used to extract translingual knowledge from texts for different purposes, in particular, MT and CLIR. We will begin with a description of the classical approaches to statistical machine translation, and describing how statistical translation models can be constructed from parallel texts, and examining extensions to the classical approaches that attempt to go beyond word-based translation. In the remaining sections, we consider a variety of methods for translingual text mining for CLIR applications.

2. Traditional Translingual Text Mining – Machine Translation

The goal of machine translation (MT) is to use a computer system to translate a text written in a source language (e.g., Chinese) into a target language (e.g., English). In this section, we provide an overview of translation models that are widely used in state-of-the-art statistical machine translation (SMT) systems. A comprehensive review is provided in a very readable form in Koehn (2009). Although these models are designed for translating regular natural language sentences, they can also be adapted to the task of search query translation for cross-lingual information retrieval (CLIR), as will be discussed in Section 4. The query translation task differs from conventional text translation mainly in its treatment of word order. In text translation word order is crucial to the readability of the translated sentences which are presented directly to the end users. In CLIR, query translation is an intermediate step that provides a set of translated query terms so that the search engine can retrieve documents in the target language. Word order thus has little impact on the search quality as long as the translation preservers the underlying search intent, rather than the form, of the original query. This section focuses only on statistical translation models for regular text. Readers who are interested in statistical models for query translation may refer to Gao and Nie (2006) and Gao et al. (2001, 2002).

2.1 SMT and Generative Translation Models

SMT is typically formulated within the framework of the noisy channel model. Given a source sentence (in Chinese) $C = c_1 \dots c_J$, we want to find the best English translation $E = e_1 \dots e_I$ among all possible translations:

$$E^* = \underset{E}{\operatorname{argmax}}\, P(E|C) \tag{10.1}$$

where the argmax operation denotes the decoder, i.e., the search algorithm used to find the target sentence with the highest probability among all possible targets.

Applying Bayes' decision rule and dropping the constant denominator, we have

$$E^* = \underset{E}{\operatorname{argmax}}\, P(C|E)P(E) \tag{10.2}$$

where $P(E)$ is the language model, assessing the overall well-formedness of the target sentence, and $P(C|E)$ is the translation model, modeling the transformation probability from E to C. In this section, we focus our discussion on the translation model only. Notice that, mathematically, the translation direction changes from $P(E|C)$ in Equation (2.1) to

$P(C|E)$ in Equation (2.2) when Bayes rule is applied. Following Koehn (2009), we will seek to avoid potential confusion that might arise from this alternation by adhering to the notation $P(C|E)$.

In a significant generalization of the noisy channel model, Och and Ney (2002) introduced a log-linear model that models $P(E|C)$ directly. This log-linear model is currently adopted by most of state-of-the-art SMT systems and is of the form

$$P(E|C) = \frac{1}{Z(C,E)} \exp \sum_i \lambda_i h_i(C,E) \tag{10.3}$$

where Z is the normalization constant, $h(\cdot)$ are a set of features computed over the translation and λ's are feature weights optimized on development data using e.g., minimum error rate training (Och 2003). The features used in the log-linear model can be binary features or real-value features derived from probabilistic models. For example, we can define the logarithm of language model and translation model probabilities in Equation (2.2) as two features, thereby subsuming the noisy channel model as a special case. The log-linear model thus provides a flexible mathematical framework with which to incorporate a wide variety of features useful for MT.

Conceptually, a translation model tries to "remember" to the extent possible how likely it is that a source sentence translates into a target sentence in training data. Figure 1 shows a Chinese sentence paired with its English translation. Ideally, if the translation model could remember such translation pairs for all possible Chinese sentences, we would have a perfect translation system. Unfortunately, a training corpus, no matter how large, can cover only a tiny fraction of all possible sentences. Given limited training data, it is usual to break the sentences in the training corpus into smaller *translation units* (e.g., words) whose distribution (i.e., translation probabilities) can be more easily modeled. In Figure 1, although the translation of the full sentence is unlikely to occur in training data, individual word translation pairs such as (rescue, 救援) will be found. Given an input sentence that is unseen in training data, an SMT system can be expected to perform a translation process that runs broadly as follows: first the input source sentence is broken into smaller translation units, then each unit is translated into a target language, and finally the translated units are glued together to form a target sentence. The translation models that we detail in the sections below differ in how the translation units are defined, translated and reassembled. The method we use to formulate a translation model is called **generative modeling**, and consists of three steps:

- Story making: formulating a generative story about how a target sentence is generated step by step from a source sentence.
- Mathematical formulation: modeling each generation step in the generative story using a probability distribution.
- Parameter estimation: implementing an effective way of estimating the probability distributions from training data.

These three modeling tasks are closely interrelated. The way in which we break the generation process into smaller steps in our story determines the complexity of the probabilistic models, which in turn determines the set of the model parameters that need to be estimated. We can view the three tasks as straddling the artistic (story making), the scientific (mathematical formulation), and the engineering (parameter estimation). The overall challenge of generative modeling is to find a harmonic combination of the three, an intellectual endeavor that attracts the talent of some of the best computer scientists all over the world.

State-of-the-art translation models used for conventional text translation broadly fall into three categories: word-based models, phrase-based models, and syntax-based models. In what follows, we will describe them in turn starting with the generative story, then describing the mathematical formulation and the way in which the model parameters are estimated on the training data.

救援(rescue) 人员(staff) 在(in) 倒塌的(collapsed) 房屋(house)
里(in) 寻找(search) 生还者(survivors)。
Rescue workers search for survivors in collapsed houses.

Figure 10.1. A Chinese sentence and its English translation

2.2 Word-Based Models

Word-based models use words as translation units. The models stem from pioneering work on statistical machine translation conducted by an IBM group in the early 1990s. In what has become classical paper Brown et al. (1993) proposed a series of word-based translation models of increasing complexity that come to be known as the IBM Models.

IBM Model 1, one of the simplest and most widely used word-based models, is what is termed a *lexical translation model*, in which the order of the words in the source and target sentence is ignored. The generative story about how the target sentence E is generated from the source sentence C, runs as follows:

1 First choose the length for the target sentence I, according to the distribution $P(I|C)$.

2 Then, for each position $i(i = 1 \ldots I)$ in the target sentence, we choose a position j in the source sentence from which to generate the i-th target word e_i according to the distribution $P(j|C)$, and generate the target word by translating c_j according to the distribution $P(e_i|c_j)$. We include in position zero of the source sentence an artificial "null word", denoted by <null>the purpose of which is to allow the insertion of additional target words.

Now, let us formulate the above story mathematically. In Step 1, we assume that the choice of the length is independent of C and I, thus we have $P(I|C) = \epsilon$, where ϵ is a small constant. In Step 2, we assume that all positions in the source sentence, including position zero for the null word, are equally likely to be chosen. Thus we have $P(j|C) = \frac{1}{J+1}$. Then the probability of generating e_i given C is the sum over all possible positions, weighted by $P(j|C)$: $P(e_i|C) = \sum_j P(j|C)P(e_i|c_j) = \frac{1}{J+1}\sum_j P(e_i|c_j)$. Assuming that each target word is generated independently from C, we end up with the final form of IBM Model 1.

$$P(E|C) = P(I|C)\prod_{i=1}^{I} P(e_i|C) \tag{10.4}$$

$$= \frac{\epsilon}{(J+1)^I}\prod_{i=1}^{I}\sum_{j=0}^{J} P(e_i|c_j) \tag{10.5}$$

We can see that IBM Model 1 has only one type of parameter to estimate, the lexical translation probabilities $P(e|c)$. If the training data consists of sentence pairs that are word-aligned as shown in Figure 2, $P(e|c)$ can be computed via Maximum Likelihood Estimation (MLE) as follows:

$$P(e|c) = \frac{N(c,e)}{\sum_{e'} N(c,e')} \tag{10.6}$$

where $N(c,e)$ is the number of times that the word pair (c,e) is aligned in training data. In practice, it is more realistic to assume that training data is aligned at the sentence level but not at the word level. Accordingly, we apply the Expectation Maximization (EM) algorithm to compute the values of $P(e|c)$ and the word alignment iteratively. This process will determine the best $P(e|c)$ that maximizes the probability of the given alignment between sentences. The algorithm works as follows:

1 Initialize the model with a uniform translation probability distribution.

2 Apply the model to the data, computing the probabilities of all possible word alignments.

3 (Re-)estimate the model by collecting counts for word translation over all possible alignments, weighted by their probabilities computed in the Step 2.

4 Iterate through Steps 2 and 3 until convergence.

Since at every EM iteration the likelihood of the model given the training data is guaranteed not to decrease, the EM algorithm is guaranteed to converge. In the case of IBM Model 1, it is guaranteed to reach a global maximum.

Brown et al. (1993) presents five word-based translation models of increasing complexity, namely IBM Model 1 through 5. In IBM Model 1 the order of the words in the source and target sentences is ignored, and the model assumes that all word alignments are equally likely. Model 2 improves on Model 1 by adding an absolute alignment model in which words that follow each other in the source language have translations that follow each other in the target language. Models 3, 4, and 5 model the "fertility" of the generation process with increasing complexity. Fertility is a notion reflecting the observation that an input word in a source language tends to produce a specific number of output words in a target language. The fertility model captures the information that some Chinese words are more likely than others to generate multiple English words. All these models have their individual generative stories and corresponding mathematical formulations, and their model parameters are estimated using the EM algorithm. Readers may refer to Brown et al. (1993) for details.

2.3 Phrase-Based Models

Phrase-based models are the basis for most state-of-the-art SMT systems. Like the word-based models, these are generative models that translate an input sentence in a source language $\boldsymbol{C}$ into a sentence in a target language E. Instead of translating single words in isolation, however, phrase-based models translate sequences of words (i.e., phrases) in C into sequences of words in E. The use of phrases as translation units is motivated by the observation that one word in a source language frequently translates into multiple words in a target language, or vice versa. Word-based models cannot handle these cases adequately: the

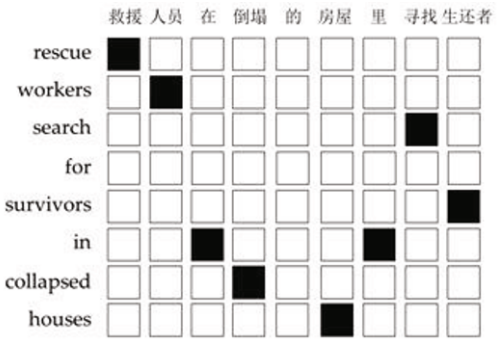

Figure 10.2. Word alignment: words in the English sentence (rows) are aligned to words in the Chinese sentence (columns) as indicated by the filled boxes in the matrix

English phrase "stuffy nose", for example, translates the Chinese word "鼻塞" with relatively high probability, but neither of the individual English words "stuffy" and "nose" has a high word translation probability to "鼻塞".

The generative story behind the phrase-based models can be stated as follows. First, the input source sentence C is segmented into K nonempty word sequences $\mathbf{c_1}, \ldots, \mathbf{c_K}$. Then each is translated to a new nonempty word sequence $\mathbf{e_1}, \ldots, \mathbf{e_K}$. Finally these phrases are permuted and concatenated to form the target sentence E. Here $\mathbf{c}$ and $\mathbf{e}$ denote consecutive sequences of words.

To formalize this generative process, let S denote the segmentation of C into K phrases $\mathbf{c}_{1\ldots}\mathbf{c}_K$, and let T denote the K translation phrases $\mathbf{e}_1 \ldots \mathbf{e}_K$ We refer to these $(\mathbf{c}_i, \mathbf{e}_i)$ pairs as *bilingual phrases*. Finally, let M denote a permutation of K elements representing the final reordering step. Figure 3 demonstrates the generative procedure.

Next let us place a probability distribution over translation pairs. Let $B(C, E)$ denote the set of S, T, M triples that translate C into E. If we assume a uniform probability over segmentations, then the phrase-based translation model can be defined as:

$$P(E|C) \propto \sum_{(S,T,M)\in B(C,E)} P(T|C,S) \cdot P(M|C,S,T) \tag{10.7}$$

It is common practice in SMT to use the maximum approximation to the sum: the maximum probability assignment can be found efficiently by using a dynamic programming approach:

$$P(E|C) \approx \max_{(S,T,M)\in B(C,E)} P(T|C,S) \cdot P(M|C,S,T) \tag{10.8}$$

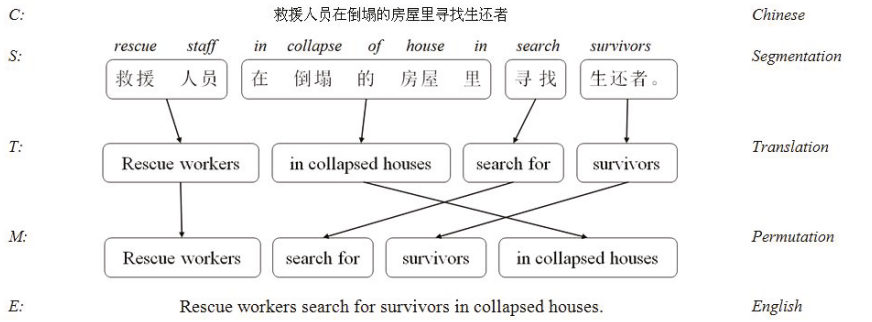

Figure 10.3. Example demonstrating the generative procedure behind the phrase-based model.

Reordering is handled by a distance-based reordering model (Koehn et al. 2003) relative to the previous phrase. We define $start_i$ as the position of the first word of the Chinese input phrase that translates to the i-th English phrase, and end_i as the position of the last word of that Chinese phrase. The reordering distance is computed as $start_i - end_i - 1$, i.e., the number of words skipped when taking foreign words out of sequence. We also assume that a phrase-segmented English sentence $T = \mathbf{e}_1 \ldots \mathbf{e}_K$ is generated from left to right by translating each phrase $\mathbf{c}_1 \ldots \mathbf{c}_K$ independently. This yields one of the best-known forms of phrase-based model:

$$P(E|C) \propto \max_{(S,T,M) \in B(C,Q)} \prod_{k=1}^{K} P(\mathbf{e_k}|\mathbf{c_k}) d(start_i - end_{i-1} - 1) \qquad (10.9)$$

In Equation (10.9) the only parameter to be estimated is the translation probabilities on the bilingual phrases $P(\mathbf{e}|\mathbf{c})$. In what follows, we rely mainly on work by Och and Ney (2002) and Koehn et al. (2003) to describe how bilingual phrases are extracted from the parallel data and $P(\mathbf{e}|\mathbf{c})$ is estimated.

First, we learn two word translation models via EM training of a word-based model (i.e., IBM Model 1 or 4) on sentence pairs in two directions: from source to target and from target to source. We then perform Viterbi word alignment in each direction according to the corresponding model for that direction. The two alignments are combined, starting with the intersection of the two alignments, and gradually including more alignment links according to heuristic rules detailed in Och and Ney (2002). Finally, bilingual phrases that are consistent with the word alignment are extracted. Consistency here implies two things. First, there must be at least one aligned word pair in the bilingual phrase. Second, there

must be no word alignments from words inside the bilingual phrase to words outside the bilingual phrase. That is, we do not extract a phrase pair if there is an alignment from within the phrase pair to an element outside the phrase pair. Figure 4 illustrates the bilingual phrases we can generate from the word-aligned sentence pair by this process.

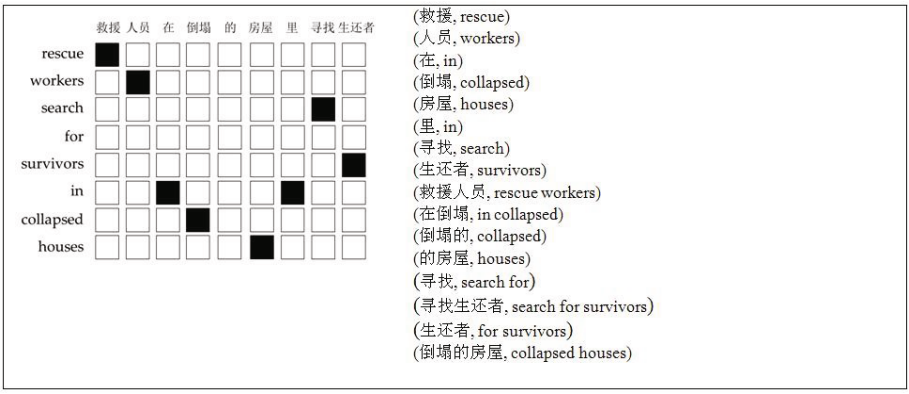

Figure 10.4. An example of a word alignment and the bilingual phrases containing up to 3 words that are consistent with the word alignment.

After gathering all such bilingual phrases from the training data, we can estimate conditional relative frequency estimates without smoothing. For example, the phrase transformation probability $P(\mathbf{e}|\mathbf{c})$ in Equation (2.7) can be estimated approximately as:

$$P(\mathbf{e}|\mathbf{c}) = \frac{N(\mathbf{c}, \mathbf{e})}{\sum_{\mathbf{e}'} N(\mathbf{c}, \mathbf{e}')} \tag{10.10}$$

where $N(\mathbf{c}, \mathbf{e})$ is the number of times that $\mathbf{c}$ is aligned to $\mathbf{e}$ in training data. These estimates are useful for contextual lexical selection when there is sufficient training data, otherwise can be subject to data sparsity issues.

An alternate means of estimating translation probabilities that is less susceptible to data sparsity is the so-called *lexical weight* estimate. Assume we have a word translation distribution $t(e|c)$ (defined over individual words, not phrases), and a word alignment A between $\mathbf{e}$ and $\mathbf{c}$; here, the word alignment contains (i, j) pairs, where $i \in 1 \ldots |\mathbf{e}|$ and $j \in 0 \ldots |\mathbf{c}|$, with 0 indicating an inserted word. Then we can use the following estimate:

$$P_w(\mathbf{e}|\mathbf{c}, A) = \prod_{i=1}^{|\mathbf{e}|} \frac{1}{|\{j|(j,i) \in A\}|} \sum_{\forall (i,j) \in A} t(e_i|c_j) \tag{10.11}$$

We assume that for every position in $\mathbf{e}$, there is either a single alignment to 0, or multiple alignments to non-zero positions in $\mathbf{c}$. In effect, this computes a product of per-word translation scores; the per-word scores are averages of all the translations for the alignment links of that word. We estimate the word translation probabilities using counts from the word aligned corpus: $t(e|c) = \frac{N(c,e)}{\sum_{e'} N(c,e')}$. Here $N(c, e)$ is the number of times that the words (not phrases as in Equation (2.8)) c and e are aligned in the training data. These word-based scores of bilingual phrases, though not as effective in contextual selection as previous ones, are more robust to noise and sparsity. Both model forms of Equation (2.8) and (2.9) are used as features in the log-linear model for SMT as Equation (2.3).

2.4 Syntax-Based Models

The possibility of incorporating syntax information in SMT has been a long-standing topic of research. Syntax-based translation models have begun to perform as well as state-of-the-art phrase-based models, and in the case of some language pairs may even outperform their phrase-based counterpart. Research on syntax-based models is a fast-moving area, with numerous open questions. Our description in this section focuses on some basic underlying principles, illustrated by examples from the most successful models proposed so far (e.g., Chiang 2005; Galley et al. 2004).

Syntax-based models rely on parsing the sentence in either the source or the target language, or in some cases in both. Figure 5 depicts the sentence pair from Figure 1, but with constituent parses added. These parses are generated from a statistical parser trained on Penn Treebank. Each parse is a rooted tree where the leaves are original words of the sentence and the internal nodes cover a contiguous sequence of the words in the sentence, called a constituent Each constituent is associated with a phrase label describing the syntactic role of the words under its node.

The tree-structured parse plays similar roles in syntax-based models to those of a phrase in phrase-based models. The first role is to identify translation units in an input sentence. While in phrase-based models the units are phrases, in syntax-based models they are constituents of the kind seen in Figure 5. The second is to guide how best to glue those translated constituents into a well-formed target sentence. Again, we assume a generative story, similar to that for phrase-based models:

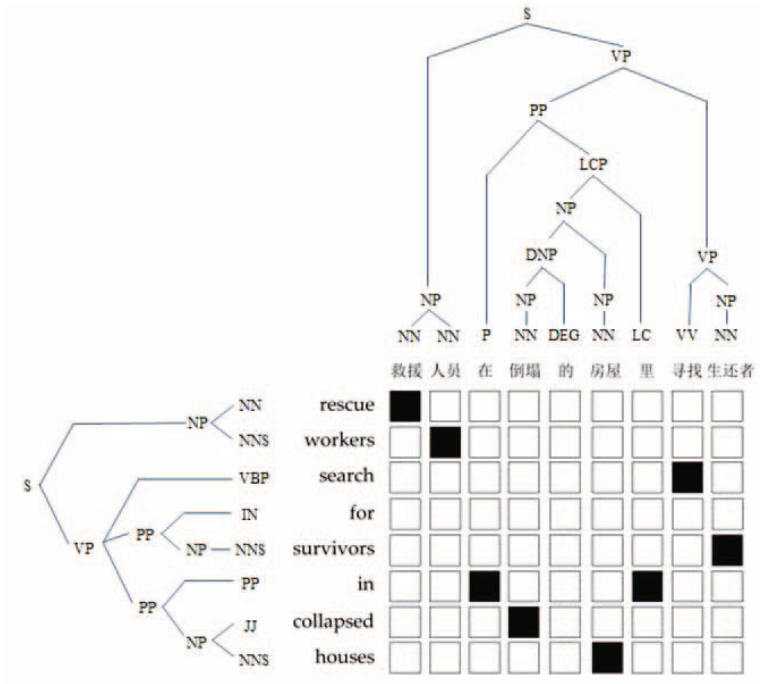

Figure 10.5. A pair of word-aligned Chinese and English sentences and their parse trees.

1 Parse an input Chinese sentence into a parse tree

2 Translate each Chinese constituent into English

3 Glue the English constituents into a well-formed English sentence.

This generative process is typically formulated under the framework of weighted synchronous Context Free Grammar (CFG) (Chiang, 2005), which consists of a set of rewriting rules r of the form:

$$X \rightarrow (\gamma, \alpha, \sim) \tag{10.12}$$

where X is a nonterminal, γ and α are both strings of terminals and non-terminals corresponding respectively to source and target strings, and ˜indicates that any non-terminals in the source and target strings are aligned. For example, a rule extracted from the example in Figure 5 is:

$$\text{VP} \rightarrow (\text{PP寻找NP, search for NP PP}, \sim) \tag{10.13}$$

where $\sim$ indicates that PP and NP in the source and target languages are aligned. We can see that these non-terminals generalize the phrases used in the phrase-based models described in Section 2.2.

We now define a derivation D as a sequence of K rewrites $r_1, \ldots, r_K$, each of which picks a rewriting rule from the grammar, and rewrites a constituent in Chinese into English, until an English sentence is generated. Let $E(D)$ be the English strings generated by D, and $C(D)$ be the Chinese strings generated by D. Assuming that the parse tree of the input Chinese sentence is $Tree(C)$, the translation model can be formulated as

$$P(E|C, Tree(C)) = \sum_{\substack{D:E(D)=E \\ \text{and } C(D)=C}} P(D) \tag{10.14}$$

As when formulating the phrase-based models, we use the maximum approximation to the sum:

$$P(E|C, Tree(C)) \propto \max_{\substack{D:E(D)=E \\ \text{and } C(D)=C}} P(D) \tag{10.15}$$

A synchronous CFG assumes that each rewriting rule application depends only on a non-terminal, and not on any surrounding context. Thus we have:

$$P(D) = \prod_{k=1}^{K} P(r_k) \tag{10.16}$$

Rewriting rule (2.11) not only specifies lexical translations but also encapsulates nicely the kind of reordering involved when translating Chinese verb complexes into English. As a result, searching for the derivation that has the maximum probability assignment, as in Equation (2.13), simultaneously accomplishes the two tasks of constituent translation and sentence reordering (as in Steps 2 and 3 in our generative story). The search can be achieved by chart parsing.

The synchronous grammar proposed in Chiang (2005) illustrates how these rewriting rules may be extracted from data and how their probabilities are estimated. The grammar has not underlying linguistic interpretation and uses only one non-terminal X. Assume that we have the word-alignment sentence pair, as shown in figure 4. First, we extract initial bi-phrases that are consistent with the word-alignment, as described in Section 2.2. We write these bilingual phrases in the form of synchronous grammar:

$X \rightarrow$ (在倒塌的房屋里寻找生还者, srch. for surviv. in collap. home, $\sim$)

We then generalize these rules by replacing some substrings with the nonterminal symbol X:

$$X \rightarrow (\text{在}X_1\text{里寻找}X_2, \text{search for } X_2 \text{ in } X_1, \sim)$$

using subscript indices to indicate which occurrences of X are linked by $\sim$. This rule captures information about both lexical translation and word reordering, with the result that the learned grammar can be viewed as a significant generalization of phrase-based models capable of handle longer range word reordering.

To limit the number of rules generated in this fashion, the rewrite rules are constrained: (a) to contain at least one and at most 5 lexical items per language, (b) to have no sequences of non-terminals, (c) to have at most two non-terminals, and (d) to span at most 15 words. Once the rewrite rules are extracted, their probabilities are estimated on the word-aligned sentence pairs using a method analogous with that for the phrase-based models. Readers may refer to Chiang (2005) for a detailed description.

3. Automatic Mining of Parallel texts

The previous section provides an overview of the state of the art in SMT. It also describes the most traditional way to exploit a parallel corpus to extract translational knowledge in form of translation models. These models are the basis for many applications in which translation is required.

SMT requires a large number of parallel texts for model training. Traditionally, one assumed that such parallel texts are available. Indeed, there have been several manually compiled large parallel corpora available. The Canadian Hansard[1] is probably the most widely used and best known. This corpus contains all the debates in the Canadian parliament in both English and French. Translation is made by professionals and it is of very high quality. The first research work on statistical MT has been carried out using this corpus. Later on, several other parallel corpora became available, in particular, the Hong Kong law documents in English and Chinese[2] and the documents of the European Parliament in several European languages[3]. These manually compiled parallel texts can be used in the methods presented in the previous section, often after a step of sentence alignment (Gale and Church 1993).

[1] http://www.parl.gc.ca/ParlBusiness.aspx?Language=E
[2] http://www.legco.gov.hk/english/index.htm
[3] http://www.europarl.europa.eu/

However, despite the high quality of translation in these parallel corpora, we do encounter several problems when they are used for the translation purposes. Indeed, although the size of the corpora are large, it is still limited for the purpose of model training, leaving a considerable proportion of the translation phenomena either uncovered or insufficiently covered for general translation applications In particular:

- **Vocabulary** The documents in these manually compiled parallel corpora are formal in style and vocabulary. They do not provide good coverage of terms or words used in less formal discussions and communications on the Web. Many terms and words in the latter will be "unknown" by the models trained on these data
- **Structure** High quality documents and their translations are written in correct syntax. This is not the case for Web documents and search queries. A syntax-based SMT system trained on these data will be inadequate to cope with the flexible structure of texts and queries on the Web.
- **Adaptability** Because the statistical translation models are trained on the parallel texts, they tend to fit the latter, including the frequency of word usage and word translation. Even if a word or a term is well covered by translation model, the suggested translations may not be suitable for the intended application.

One possible solution to the above problems is to develop automatic tools to collect appropriate parallel documents according to one's requirement. The Web is an excellent resource for this purpose, and indeed it is a truly multilingual resource, one on which documents in many different languages are published. A certain proportion of the documents are parallel, i.e. the same documents are published in several languages. These documents virtually constitute a large parallel corpus. The key problem is to collect those parallel texts without including (too many) non-parallel ones.

Attempts to collect parallel texts from the Web date to the late 1990s, with Resnik (1998) and Nie et al. (1999). Both studies exploit two factors to determine whether two texts are parallel: the Web structure in which the texts are stored and published, the text structure of the documents themselves.

3.1 Using Web structure

Resnik (1998) observed that in many cases, parallel Web pages are linked from an entry page (home page) on a website, each with a language

identifier such as “English” and “Français” as anchor text. For example, the following website (Natural Resources of Canada) is organized in the manner shown in Figure 6.

Figure 10.6. An example of parallel pages linked from a home page

The mining system STRAND (Resnik, 1998) identifies the referred pages as candidate parallel web pages. In the query language of Alta Vista used by STRAND, the following query will retrieve parent pages referring to two child pages in the relevant languages:

anchor:”english” AND anchor:”français”

However, the above criterion can only detect a limited number of parallel Web pages. More commonly sites are organized so that each of the pages contains a link to the corresponding parallel page, as shown in Figure 7. Again, the link usually contains anchor text that identifies the language.

To retrieve those pages, the following Alta Vista query can be used to retrieve the French documents containing an anchor text to an English page:

anchor: ”english” OR anchor: ”anglais”

while setting the language of the documents to French. Analogously, one can retrieve English documents containing anchor text linking to a French page.

Figure 10.7. An example of mutually linked parallel pages.

This second criterion is the main approach taken by PTMiner (Nie et al. 1999, Chen and Nie 2000) to identify candidate pages. PTMiner additionally used a site crawler to download all the pages from candidate sites (the sites that contain some candidate parallel pages) in order to find more Web pages on those sites that are not indexed by the search engine.

3.2 Matching parallel pages

Once two sets of candidate pages are determined, the next task is to pair the pages up. The contents of the pages will eventually be used, but first heuristics are applied to quickly identify candidate parallel pages Since parallel Web pages are usually assigned similar file names two Web pages with the names "description_en.htm" and "description_fr.htm" are likely parallel. Similarly, Web sites may use two separate directories to store pages in two languages, in which case, the names of the directories may be slightly different, e.g. "www.website.com/English/file1.html" vs. "www.website.com/French/file1.html". In both cases the difference between file and directory names is often related to the language, and this can be recognized using simple heuristics. Such heuristics are used in PTMiner to pair up mined candidate Web pages efficiently.

To further filter out non-parallel pages, additional checks on the pages' contents can then be applied:

- Are the HTML structures of the two pages similar? The assumption is that parallel pages are usually created with the same or similar HTML structures. In both STRAND and PTMiner, the HTML markup sequence of each page is extracted, and the pages are considered to be parallel if their HTML markup sequences resemble each other However, more sophisticated comparison of document structure can be performed. For example, one can use the DOM tree of the Web page (Shi et al. 2006).
- Are the two pages of similar lengths? It is generally observed that the lengths of parallel texts are similar (or proportional to the length ratio of texts in the two languages). This is an easy way to filter out candidate pairs whose content cannot be parallel.
- Finally, what is the content translation probability? If the texts in the two pages have a high mutual translation probability, then the pages are likely to be parallel. Although an effective means of confirming textual parallelism, this ultimate step is costly to implement and has not been widely used.

The precision of the Web pages identification by STRAND and PTMiner is impressive: it is estimated that more than 90% of the identified pairs of Web pages are indeed parallel. Evaluation of recall, on the other hand, presents greater difficulty: Resnik calculated recall of STRAND at 62.5%, while Nie et al. estimated the lower bound of recall of PTMiner at a little over 50% on the assumption that every Web page from a candidate website has a parallel page in another language, an assumption that obviously overestimates the case Nevertheless, lower recall ratios can be tolerated because the number of potential parallel pages on the Web is very large, and it is more important to have a mining process with high precision than high recall.

In term of volume, STRAND has mined a relatively small number of parallel pages while PTMiner has successfully collected large amounts of parallel page data in English-French, English-Chinese, English-German, etc., chiefly by exploiting criteria that correspond to more commonly employed techniques of organizing parallel pages on the Web, as well as the site crawling process.

The above mining strategy has been used in a number of studies pertaining to different language pairs: Ma and Liberman (1999) used a similar approach to PTMiner to mine parallel pages in German and English, with some slight differences in the process: the similarity between

the file names of candidate pages is measured by edit distance, and the known translations are mapped with some position constraint within the texts. Similar approaches have been used by Nagata et al. (2001) and Yang and Li (2003) to mine English-Japanese and English-Chinese parallel pages. Resnik et al. (2003) have further explored the mining of parallel Web pages from Web archives.

The above processes are designed for mining on general websites. It is possible to incorporate additional criteria according to the specific organization of a website. For example, parallel texts on the same website (e.g. Wikipedia, newswire websites) can share common resources such as pictures. Metadata can also be incorporated in documents. The use of such indications can further improve the mining process.

Bilingual and multilingual newswire websites are a common source from which parallel texts are mined. Many newswire publishers publish articles in several languages, and in many cases, the articles in different languages are translations. For example, China Daily publishes certain bilingual news articles that are aligned in paragraphs. Some of the news articles are translated and published in several languages such as Chinese, English and French. Several European newspapers also publish simultaneously articles in several languages. This provides an easy way to collect parallel news articles. However, the collection of parallel news articles depends on the specific organization of each newspaper. In some cases, there is a systematic schema of correspondence, while in other cases no clear structural information is available to determine whether two articles are parallel. In the latter case, the mined result is often comparable texts rather than parallel texts. We will describe some attempts of this kind in Section 5.

4. Using Translation Models in CLIR

It is safe to assume that not all automatically mined Web pages are strictly parallel. Indeed, during manual evaluation, it turns out that some pages, which presumably should contain the same information, are not parallel in content: one of the pages can be outdated, contain only part of the information, or even consist of an "under construction" message. The precision and recall numbers mentioned earlier are subject to human judgment: If the contents are parallel above some threshold, we consider the pages to be parallel a situation that is less ideal than the Hansard corpus, especially for tasks such as full-text machine translation that call for high quality parallel texts for training.

For other less demanding tasks such as CLIR, however, translation models trained from automatically collected Web pages can perform very

well. A translation model trained on a parallel corpus can be naturally integrated into the CLIR process. General IR can be processed using a language modeling approach as follows:

$$Score(Q, D) = \sum_{t \in V} P(t|M_Q) \log P(t|M_D) \tag{10.17}$$

where M_Q and M_D are respectively a statistical language model estimated for the query and the document, and V is the vocabulary. Notice that both M_Q and M_D are generation models, i.e. no word order or relationship is taken into account. Such an approach is often called "bag of words" approach. Using such an approach, translation in CLIR is also performed at the word level: Each word is translated independently. Therefore, the simple IBM Model 1 is widely used.

For CLIR, either the document or the query should be translated. One can of course use an MT system to translate them. Because the Web search engine only uses words and ignores word order, MT offers more than what is needed. One may argue that this is not necessarily a bad thing to have a tool offering more than required. Indeed, in the CLIR experiments, it is usually found that a high-quality MT system leads to a good CLIR result when it is used to translate queries or documents.

However, off-the-shelf MT systems also have weaknesses:

- An MT system chooses only one translation word (or expression) for each source word. In reality, there may be multiple translations. For example, "drug" (illegal substance) can be translated into "drogue" or "stupéfiant" in French. By limiting to one translation, documents in French using the other term cannot be found. For CLIR, keeping multiple translations for a word is often preferred.

- The translation by an MT system is limited to the true "translations" of the words in a query or a document. In IR, on the other hand, it is usually preferred to add related terms in the query (or document) to expand it. Query (or document) expansion is a common method in IR to increase retrieval effectiveness. By including only true translation words in a query translation, CLIR does not benefit from query expansion. It is preferred to include also related terms in the target language when doing query translation in CLIR.

- The final translation result by an MT system does not distinguish the words in their importance, i.e. all the words are un-weighted.

In IR, the weighting of terms in the query is crucial, and the translation probability or weight can greatly help distinguishing important terms vs. unimportant ones.

The above reasons have motivated a number of attempts to design CLIR approaches without, or in addition to, the use of MT systems. The principle of CLIR can be well described within the language-modeling framework for IR. It includes a translation of the query or document model as follows:

$$Score(Q, D) = \sum_{t \in V_t} [\sum_{s \in V_s} P(t|s)P(s|M_Q)] \log P(t|M_D) \tag{10.18}$$

$$Score(Q, D) = \sum_{t \in V_t} P(s|M_Q) \log [\sum_{t \in V_t} P(s|t)P(t|M_D)] \tag{10.19}$$

in which V_s and V_t are respectively the vocabulary in source and target languages, and $P(s|t)$ and $P(t|s)$ are translation probability (in IBM model 1) of a target language term (t) to a source language term (s) and vice versa. In practice, rather than using the whole vocabulary in $\sum_{s \in V_s} P(t|s)P(s|M_Q)$ and $\sum_{t \in V_t} P(s|t)P(t|M_D)$, one can select a subset of the translation terms, for example, the translation terms whose translation probability is higher than a threshold, or the N best translation terms for the query. Different from general MT, query or document translation in CLIR usually selects multiple translation words (rather than the best one), thereby producing a desired expansion effect. In addition, the translation probability is used explicitly to determine term weighting for the retrieval process.

The use of automatically mined parallel corpora in CLIR has been successful. In an early experiment on CLIR, Nie et al. (1999) reported that using the Web corpus, the CLIR effectiveness is very similar to using the Hansard corpus. Further experiments (Kraaij et al. 2003) have shown that CLIR using the Web parallel corpus outperforms methods that use an existing dictionary-based MT system - Systran. These results indicate that CLIR does not require as high quality corpora for training translation models. A noisy corpus can be as effective as a manually compiled high-quality corpus. In addition, a query or document translation properly incorporated into the retrieval model (as Equations 2.16 and 2.17) is a better solution than using an MT system as an individual tool, separated from the retrieval model.

5. Collecting and Exploiting Comparable Texts

The success of using a noisy parallel corpus in CLIR indicates that one can tolerate certain noise in the text data used for model training. To what extent is the process tolerant to noise? There is no clear answer to this question, but there is a series of experiments using comparable texts for CLIR, which have shown encouraging results: comparable texts are good complements to other translation resources.

In general, comparable texts are defined as texts that are not necessarily parallel, but describe the same event. Other terminologies are also used. Fung and Cheung (2004) defined quasi-comparable and comparable documents because they were written independently but on more or less the same topic. Noisy-parallel documents refer to a pair of source and translated documents that were either adapted or evolved in different ways such as Wikipedia articles. There are indeed a variety of comparable texts with different degrees of relatedness. Fung (1995) considers a continuum from parallel, comparable to unrelated texts. Brashchler and Schaüble (1998) defined the following levels of relatedness:

1 Same story: The two documents deal with the same event.

2 Related story: The two documents deal with the same event or topic from slightly different viewpoints. Or one of them deals with the topic from a broader story.

3 Shared aspect: The documents deal with related events. They may share locations or persons.

4 Common terminology: The events or topics are not directly related, but the documents share a considerable amount of terminology.

5 Unrelated: The similarities between the documents are slight or nonexistent.

Depending on the process used, different types of comparable texts can be collected. In general, the following indicators can be used to determine comparable texts from a website, especially from a newswire:

- The publication dates of two comparable texts should be the same or close;
- Some articles incorporate metadata to describe the content categories, in which case, the category of the comparable texts should be the same;

- The fact that two texts contain links to the same objects (e.g. pictures) increases the chance that the texts are about the same event;

- Although one cannot expect exact mutual translation at sentence level between comparable texts, the main vocabulary should be translatable and this can be verified using a simple resource such as a bilingual dictionary;

- The texts may contain similar special elements: named entities – they talk about the same persons and describe events of the same dates, or domain-specific words and their translations;

- Using a CLIR method, one can form a query with a source language text, and retrieve a set of potential comparable texts in the target language.

Sheridan and Ballerini (1996) are among the first to exploit comparable texts for CLIR. They mined newspaper articles in German and Italian from the website of *Schweizerische Depeschenagentur* (SDA) using content descriptor metadata and publication dates.

In their study, Brashchler and Schaüble (1998) "translated" the named entities as well as words from the source language text, and used the translation to retrieve comparable texts. Their evaluation revealed that about 60% of the texts mined are documents that share one or more events, and 75% of them share a common terminology. The mined texts have been used in a CLIR task, leading to a retrieval result only slightly worse than the best participants in TREC-7. Similar approaches have been used in other studies (e.g. Talvensaari et al, 2006, Talvensaari 2007, Huang et al. 2010). Huang et al (2010) investigated the translation of key terms in the above process: Not only single-word terms but also multi-word terms are extracted from the source-language document and translated. By doing so, they reduced the translation ambiguity, and produced more precise description in the target language.

Instead of using the translated terms in a CLIR process to mine comparable texts, in several studies, the frequencies and ranks of the source terms and their translations have also been used. Fung and Lo (1998), Fung and Cheng (2004) and Carpuat et al. (2006) used a different approach to align comparable texts. They use a set of seed words, for which the translations are known. Seed words in source- and target-language texts are extracted and their frequencies are compared. It is assumed that the seed words should be comparable in their frequency ranks. Tao et al. (2005) used a more elaborated method based on Pearson correla-

tion: words and their known translations in a pair of comparable texts should have a strong correlation in their ranks.

The mined comparable texts can be used to derive a general bilingual lexicon (Rapp, 1995) or for translations of specific named entities (Fung, 1995, Ji, 2009). In general, it is more difficult to train a translation model using comparable texts than using parallel texts. A less strict bilingual term similarity is determined instead. The principle is analogous to word co-occurrence analysis in monolingual texts: two terms in different languages have a strong translingual relationship if they co-occur often in comparable texts in respective languages. The following formula (or some variants) can be used:

$$sim(w_s, w_t) = \frac{coocc(w_s, w_t)}{Z} \tag{10.20}$$

where w_s and w_t are source and target words, $coocc(w_s, w_t)$ is a measure of their co-occurrence and Z a normalization factor. $coocc(w_s, w_t)$ can take different forms: the number of pairs of comparable documents which contain the two words respectively, the minimal frequency of the two terms in the respective document, or some transformed measure based on these. As not all the words in the source document have their translations in the target document, the translingual relationships can be built up only for the most frequent words, or for named entities (Fung, 1995, Ji, 2009). Needless to say, the translingual relationships are much less precise than those extracted from truly parallel texts. There are two main reasons:

- The comparable texts are noisier by nature. A pair of comparable documents is not mutual translation, and the relationships between terms extracted from them are more translingual related than translation relations.
- As no process similar to sentence alignment on parallel texts can be performed, it is usually assumed that a word in a document corresponds to any word in the document in another language. In other words, the correspondence is not bound within a smaller portion of text than the entire document. The translingual relationships extracted are very noisy.

The translingual relationships can be hardly used alone for MT. At best, it can be used to complement other translation resources. For CLIR, the noisy translingual relationships extracted from comparable corpora have been found to perform quite well (Braschler and Shaüble, 1998) indicating that the utility of comparable texts, when exploited in a simple manner, is limited to less demanding tasks such as CLIR.

An alternative approach to exploiting a parallel/comparable corpus is pseudo-relevance feedback (Carbonell et al. 1997): Use a query in the source language to retrieve a set of texts in the parallel/comparable corpus. One can then select the set of corresponding texts in the target language, from which a set of terms can be extracted. These latter constitute a "translation" of the original query. As one may notice, this approach is similar to those on translingual term similarity. However, the difference is that, rather than determining the translingual relations between individual terms, this approach determines a translingual relation between sets of terms. There is potentially a larger effect of local context (Xu and Croft 1996).

Another approach is to construct a new representation space to which terms in both languages can be mapped. CLIR using Latent Semantic Indexing (Dumais et al. 1997) exploits this principle: parallel (comparable) texts are concatenated for form a composed document; A latent representation space is created and implicit translation is generated by mapping a term, a document or a query into the new space. One can also use a generative topic model instead of LSI.

In addition to the above methods, comparable texts can be exploited in a more refined manner by extracting a subset of strongly comparable or parallel parts (sentences) from them. We will describe these approaches in the next section.

6. Selecting Parallel Sentences, Phrases and Translation Words

The mining approaches described in the previous sections all rely on heuristics relating to the organization and other characteristics of parallel Web pages. Since some of the mining results are likely to non-parallel, or only partially parallel it is pertinent to ask whether it is possible and beneficial to clean the mined results in order to minimize noise.

There have been a number of attempts to extract a subset of high quality parallel texts or sentences from a corpus that has been initially mined by some other means. An original corpus can be extracted by an application such as PTMiner or STRAND. Or it might take the form a set of comparable texts minded from a newswire Web site. Even with the truly parallel corpora, a certain filtering is made. In fact, before translation models are trained on a set of parallel texts, the sentences in the texts are aligned (Gale and Church 1993) Different patterns of sentences alignment can be recognized: 0-1 or 1-0 (i.e. a sentence is aligned with no sentence), 1-1 (one sentence is aligned with one sentence), 1-2 or 2-1, and so forth. It has been observed that errors (i.e. non-parallel

sentences) most often appear in alignments other than 1-1. For example, 1-0 or 0-1 alignments may be due to insertion and deletion during the manual translation. Therefore, a simple filtering process is to use only 1-1 aligned sentence pairs for model training.

It is also possible to clean up an initial parallel corpus using other heuristics. In Nie and Cai (2001), the following criteria are used to filter the data extracted by PTMiner:

- The length ratio of the text pair should be close to the standard length ratio of the two languages;
- The proportion of the 1-1 alignments of a text pair should be high;
- A relatively large percentage of the terms should be translatable into terms of another text using a dictionary.

Any text pair that does not comply with these conditions is removed from the corpus. The experiments of Nie and Cai show that a combination of the above criteria can effectively remove some non-parallel texts and retrain the parallel ones. They also observe that translation models trained on the resulting cleaned corpus mined by PTMiner are of higher quality, and are more effective when used in CLIR.

While Nie and Cai's study sought to filter out non-parallel documents from the corpus, other researchers have attempted to extract parallel sentences more directly from comparable corpora. Munteanu and Marcu (2005) use the following process to extract parallel sentences in Chinese, Arabic, and English: 1). Candidate document pairs are first selected using their publication dates (within a date window of 5 days). 2). Candidate sentence pairs from the paired documents are selected using criteria similar to those used by Nie and Cai (2001), i.e. sentence length ratio and percentage of terms that can be translated in another sentence using a dictionary. 3). Finally, a maximum entropy classifier is used to determine if the candidate sentence pair is likely to be parallel. Similar methods have also been taken by Zhao and Vogel (2002), Utiyama and Isahara (2003) and Hong et al. (2010), who estimate sentence similarity variously on the basis of sentence length ratio, sentence alignment, IBM-1 translation model and percentage of known translations using a dictionary. In manual evaluation, it has been found that the selected sentence pairs can have a precision of 90% (Utiyama and Isahara 2003). These studies demonstrate that selecting a set of parallel sentences from a comparable corpus is possible. The experiments also showed that the extracted parallel sentences are useful for MT in some context: SMT systems that use the selected sentence pairs in combination with an initial set of parallel texts generally produce a higher BLEU score in SMT

experiments. However, when these parallel sentences are used alone, the performance is usually lower than that of using truly parallel texts.

Several studies use an iterative process to gradually select parallel sentences from a noisy corpus for model training. Fung and Leung (2004) first use a bilingual lexicon to select comparable texts and parallel sentences from the original set of documents. The selected parallel sentences are used to train a translation model, which is then used to complement the bilingual lexicon in a second round of document and sentence selection. Fung and Leung reported a precision of 67% in the extraction of parallel sentences using the adaptive method, 24% higher than a baseline method that only used a bilingual lexicon.

The common observation that word-based translation is too ambiguous for precise translation led researchers to propose phrase-based models (Ballesteros and Croft, 1997; Gao et al., 2001; Gao et al., 2006; Koehn et al. 2003). In the translingual relation mining task, likewise, one can go a step further Munteanu and Marcu (2006) word-align pairs of candidate sentences using IBM Model 1 in conjunction with additional heuristics, and treat a sequence of source language words (phrase) as parallel to a sequence of target language words if they have a strong mutual alignment score This principle is analogous to the case of phrase-based SMT (Koehn et al. 2003), where a sequence of words is considered to form a phrase if the constituent words are translated into a sequence of consecutive words in another language (see Section 2.2).

Again, the resulting translation model can be filtered so as to remove noise. One may choose to use only those translations whose probability is higher than a threshold (e.g. 0.01), or the N best translations for each word. One can also select translation terms according to the context, i.e. the query to be translated. One criterion that has been used in Gao et al. (2001; 2002; 2006) and Liu et al. (2005) is to assume that the resulting set of translation terms for a query should be consistent, i.e. they should co-occur often in the target language. Application of this criterion can remove unlikely translation terms that are inconsistent with the other words (or their translations) in the query. Interested readers can refer to these papers for details.

7. Mining Translingual Relations From Monolingual Texts

Translingual knowledge is by no means confined in texts in two different languages. It is by no means rare that one can find rich translingual knowledge within a "monolingual text", or more precisely, a mostly monolingual text that contains glosses (translations or transliterations)

inlined in the text. For example, the following is a short text in Chinese, with personal names glossed in English.

斯科特·霍夫曼(Scott Huffman)和史蒂夫·常(Steve Chang)继续解读移动搜索。

Even if one does not understand the whole Chinese sentence, it is possible to guess that 斯科特·霍夫曼 is the Chinese transliteration of "Scott Huffman" and 史蒂夫·常 the transliteration of "Steve Chang". This phenomenon frequently appears in many (especially Asian) languages, in particular, when a personal name or technical term calls for a transliteration or translation gloss. Between languages written in the same script, glossing of named entities may not be necessary. Indeed, it is seldom necessary for a personal name to be transliterated from one European language into another. However, when languages are written in completely different scripts, transliteration or translation is usually necessary.

Since our present focus is on mining Web data, we will not discuss the mechanics of transliterating personal names. In general, rules or statistical translation models trained on a set of name translations are used to determine possible correspondences between phonemes in two languages and between characters/syllables and phonemes. Interested readers may refer to (Chen et al. 2006; Jeong et al. 1999; Kuo et al. 2006; Lam et al. 2007; Qu et al. 2003; Sproat et al. 2005) for details.

The huge volume of documents on the web containing glosses of the kind seen provides us with a rich resource for mining translingual knowledge for personal and organizational names and technical terms. One common approach is to manually define a set of common patterns of glossing. Zhang and Vines (2004) identified the following patterns in a monolingual text (identified here as the target language):

... translation (source_term) ... e.g. 斯科特· 霍夫曼(Scott Huffman)
... translation, source_term ... e.g. 美国花旗银行, Citibank, ...
... translation, or source_term ... e.g. 潜在语义信息检索模型, 或LSI...

These patterns reflect the common ways of specifying the corresponding terms (or their glosses) in their original language, especially when for names of persons and organizations and for technical terms.

A typical mining process based on manually defined patterns runs as follows (Zhang and Vine, 2004): First, given a source language term (English) for which translations are sought, the term is used as a search query to retrieve Chinese (target language) documents. Then the patterns are applied to the snippets of the returned results to identify the

candidate translations. Further analysis of the candidates allows selection of the most frequent candidates. A number of studies have used the strategy (Cheng et al., 2004; Cao et al. 2007) to mine large numbers of translation relations from monolingual texts on the Web.

Additional mining criteria can be added to retrieve more relevant candidate snippets. For example, Zhang et al. (2005) and Huang et al. (2005) add related target language terms to the search query for snippets: To find a transliteration of "Leo Tolstoy" in Chinese (列夫.托尔斯泰), if one knows that the work "War and Peace" is closely connected to the author's name, then the Chinese terms "战争" (war) and "和平" (peace) can be added into the search query to locate highly related snippets.

Rather than exploiting a set of patterns to mine translingual relationships, Cheng et al. (2004) tries to mine related terms directly from the snippets returned by the search engine. Once a set of snippets is collected, a similarity measure is used to select terms that are related to the original term. Figure 10.8 shows an example using the query "yahoo" to retrieve documents in Chinese:

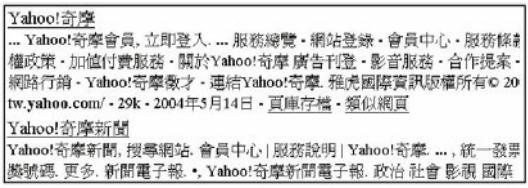
Yahoo!奇摩
... Yahoo!奇摩會員, 立即登入. ... 服務總覽 - 網站登錄 - 會員中心 - 服務條
權政策 - 加值付費服務 - 關於Yahoo!奇摩 廣告刊登 - 影音服務 - 合作提案 -
網路行銷 - Yahoo!奇摩徵才 - 連結Yahoo!奇摩. 雅虎國際資訊版權所有© 20
tw.yahoo.com/ - 29k - 2004年5月14日 - 頁庫存檔 - 類似網頁
Yahoo!奇摩新聞
Yahoo!奇摩新聞, 搜尋網站, 會員中心 | 服務說明 | Yahoo!奇摩. ... , 統一發票
獎號碼. 更多. 新聞電子報. •. Yahoo!奇摩新聞電子報. 政治 社會 影視 國際

Figure 10.8. Results of search for Chinese documents using "Yahoo" as query. (from (Cheng et al., 2004))

The snippet results contain Chinese terms strongly correlated with "yahoo" such as 奇摩(Yahoo!'s name in Taiwan) and 搜索(search). Unsurprisingly, the extracted terms are more often related terms than translations, so they may not be appropriate for use in full-text translation, but appropriate for less demanding applications such as CLIR (Cheng et al 2004). The experiments on CLIR show that these glosses supplement existing dictionaries, and can reduce the number of unknown words in query translation. This mining approach can also find additional good translations for terms that are already covered by an existing resource.

8. Mining using hyperlinks

Modern search engines often view an anchor text linking to a Web page as an alternative description of the page. When different anchor texts link to the same Web page, those anchor texts can be considered strongly

related. If the anchor texts are in different languages, moreover, then this relationship constitutes a kind of translingual/translational relationship. Figure 10.9, below, shows anchor texts in different languages pointing to the same Web page (www.yahoo.com).

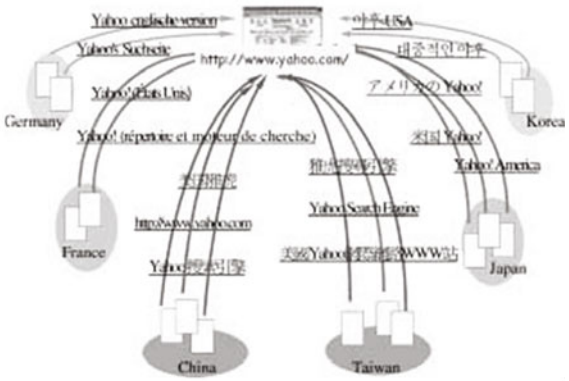

Figure 10.9. Possible hyperlinks and anchor texts to the web page www.yahoo.com. (from (Lu et al., 2004))

This is the principle used in (Lu et al. 2004) to extract translations using anchor texts. The terms "雅虎搜索引擎", "美国雅虎", "yahoo!", "Yahoo!モバイル", "Yahooの検索エンジン", etc. correspond to different names for "Yahoo!" in different languages. Lu et al. (2004) proposed a translingual similarity measure to determine relationships between terms in different languages. This approach is particularly suited to mining translations or transliterations of proper names (names of organizations and companies). It will find, for example, different transliterations of "Sony" in simplified Chinese "索尼" and in traditional Chinese "新力"; and translations and transliterations of "General Electric" or "GE" in simplified Chinese "通用电气" and in Traditional Chinese "奇异" (transliteration of "GE").

Mining on Wikipedia is a special case of hyperlink mining. Wikipedia is increasingly used in CLIR experiments to find equivalent expressions across languages, in particular proper names and technical terms. The encyclopedia contains numerous explicit links between different entries of the same entity in different languages that can be assumed to be mutual translations (Gamallo et al. 2010) For example, "Chang Kai-Shek", "Jiang Jieshi", 蒋中正and蒋介石are the different names of the same person, and they refer to the same page on Wikipedia. While the coverage provided by this resource is limited, one can further extend the mining process by also assuming that articles on the same topic in different languages are either "parallel" or comparable. These characteristics of

Wikipedia have been successfully exploited to extract translingual relations between elements in the two texts and used for CLIR (Potthast et al. 2008; Schönhofen et al. 2007; Smith et al. 2010).

9. Conclusions and Discussions

Translation is an essential component of MT and CLIR Since manually constructed resources are limited in coverage there is an acute need to acquire translingual knowledge automatically. In this chapter, we have presented a broad overview of a growing body of work on mining parallel texts, parallel sentences and phrases on the Web. These studies show that the mining processes that employ heuristics based on the organization of parallel texts and the characteristics of parallel sentences, or translation knowledge already available (e.g. a bilingual dictionary), make it possible to harvest a large amount of parallel and comparable texts on the Web. The mined texts, without cleaning, can be too noisy for tasks such as MT. However, for tasks such as CLIR, which does not always demand high-quality text translations, parallel/comparable corpora mined using these mining approaches can be directly used to train models or learn term similarity measures for query translation. Experimental results show that one can obtain improved CLIR effectiveness compared with other resources such as MT and bilingual dictionaries.

For more demanding tasks like conventional text MT, refinements can be implemented to acquire more precise translation knowledge, including filtering of the mined corpus itself, and selection of parallel sentences or parallel phrases from the corpus. Experiments with SMT models indicate that the smaller and cleaner corpora obtained by filtering do in fact help improve the translation quality in terms of BLEU score and other metrics.

Although feasibility and utility of mining translingual knowledge on the Web is now well established, much room remains for methodological improvement. Despite application of filtering techniques, a significant percentage of the mined corpora still contain non-parallel data. Such corpora may be unreliable when used to train sophisticated translation models beyond the IBM-1 models employed in most CLIR studies. For MT purposes, moreover, it may be necessary to further refine the mining process itself in order to locate strictly parallel texts and sentences. On the other end of the spectrum, although a comparable corpus is considered too noisy to be suited to translation model training approaches to smoothing the models trained using strictly parallel texts and the ones using translingual term similarity with less strictly matched texts might be applicable to produce useful models.

While it is preferable to extract well-formed phrases for general MT tasks, the requirements for other tasks such as CLIR may be less stringent. A more flexible phrase-based query translation model may well be applicable, in which, for example, context is provided by pairs of query terms, with one word defining a context for the translation of the other even though the two words themselves may not form a single phrase.

Parallel texts are essential to translation, and identifying translingual resources remains a primary goal of mining parallel texts on the Web. But parallelism need not be viewed as limited to cases involving different languages. Other kinds of data can also potentially be regarded as parallel. For example, two sets of texts in the same language can be treated as parallel and used to train a "translation" model to capture the relationships between elements in that language, an approach that has been successfully used in monolingual IR (Burger and Lafferty, 1999; Gao et al. 2010). This notion can be further extended to mining trans-media knowledge: correspondences between images and textual annotations can be exploited to generate trans-media relations between visual features and words (Jeon et al. 2003; Oumohmed et al. 2005). These studies demonstrate that the SMT paradigm is applicable in tasks other than translation and hint at the possibility of interesting new approaches in other areas.

References

[1] Adafre, S.F. and de Rijke, M. (2006). Finding similar sentences acorss multiple languages in Wikipedia. *11^{th} Conference of the European Chapter of the Association for Computational Linguistics*, pp. 62–69.

[2] Ballesteros, L. and Croft, W. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of SIGIR Conf.* pp. 84-91.

[3] Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR Conf.*, pp. 222-229.

[4] Braschler, M., and Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pp. 183–197.

[5] Braschler, M., and Schäuble, P. (2001). Experiments with the Eurospider Retrieval System for CLEF 2000, in *Proceedings of CLEF Conference.* pp. 140-148.

[6] Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), pp. 263-311.

[7] Cao, G., Gao, J., Nie, J.Y. (2007) A system to mine large-scale bilingual dictionaries from monolingual Web pages, *MT Summit*, pp. 57-64.

[8] Carbonell, J.G, Yang, Y, Frederking, R.E., Brown, R., Geng, Y. and Lee, D. (1997) Translingual information retrieval: A comparative evaluation. In: Proceedings of the International Joint Conference on Arti?cial Intelligence (IJCAI '97).

[9] Chiang, D., (2005) A Hierarchical Phrase-Based Model for Statistical Machine Translation. *ACL*.

[10] Chen, J., Nie, J.Y., (2000) Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. *ANLP pp.* 21-28

[11] Chen, H.H., Lin, W.C. and Yang, C.H. (2006). Translation-Transliterating Named Entities for Multilingual Information Access. *Journal of the American Society for Information Science and Technology*, 57(5):645-659

[12] Cheng, P., Teng, J., Chen, R., Wang, J., Lu, W., and Chien, L. (2004). Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In *Proceedings of SIGIR Conf.*, pp.162-169.

[13] Dumais, S. T., Letsche, T. A., Littman, M. L. and Landauer, T. K. (1997) Automatic cross-language retrieval using Latent Semantic Indexing. *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*, March 1997.

[14] Franz, M., McCarley, J.S. and Koukos, S. (1999) Ad hoc and multilingual information retrieval at IBM. Proceedings of the Seventh Text Retrieval Conference (TREC-7), pp. 157–168.

[15] Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. *Proceedings of the Association for Computational Linguistics*, pp. 236-243.

[16] Pascale Fung and Yuen Yee Lo. 1998. An IR approach for translating new words from nonparallel, comparable texts. *Proceedings of COLING-ACL98*, pp. 414– 420.

[17] Fung, P. and McKeown, K. (1997) Finding terminology translations from non-parallel corpora. In: The 5th Annual Workshop on Very Large Corpora.

[18] Fung, P. and Cheung, P. (2004) Multilevel boot-strapping for extracting parallel sentences from a quasi parallel corpus. *Conference on Empirical Methods in Natural Language Processing (EMNLP 04),* pp. 1051–1057.

[19] Gale, W. A., Church K. W. 1993. *A Program for Aligning Sentences in Bilingual Corpora*. Computational Linguistics, 19(3): 75-102.

[20] Galley, M., Hopkins, M., Knight, K., Marcu, D., (2004) What's in a translation rule? *HLT-NAACL*, pp. 273-280

[21] Pablo Gamallo Otero, Isaac Gonzalez Lopez, (2009) Wikipedia as Multilingual Source of Comparable Corpora, *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC 2010, pp. 21–25

[22] Gao, J., Nie, J.Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of SIGIR Conf.*, pp. 96-104.

[23] Gao, J., Zhou, M., Nie, J.Y., He, H., Chen, W. (2002) Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. *SIGIR*, pp. 183-190

[24] Gao, J., Nie, J.Y. (2006) Study of Statistical Models for Query Translation: Finding a Good Unit of Translation. *SIGIR*, pp 194-201, 2006.

[25] Gao, J., He, X., Nie. J.Y. (2010) Clickthrough-based translation models for web search: from word models to phrase models. *CIKM*, pp 1139-1148, 2010.

[26] Hong, Gumwon, Li, Chi-Ho, Zhou, Ming and Rim, Hae-Chang (2010) An Empirical Study on Web Mining of Parallel Data, *COLING*, pp. 474–482.

[27] Huang, Degen, Zhao, Lian, Li, Lishuang Yu, Haitao (2010) Mining Large-scale Comparable Corpora from Chinese-English News Collections, *COLING*, pp. 472-480.

[28] Huang, F., Zhang, Y., and Vogel, S. (2005). Mining Key Phrase Translations from Web Corpora. In *Proceedings of HLT-EMNLP Conf.,* pp. 483-490.

[29] Jeon, J. Lavrenko, V. and Manmatha, R. (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, *SIGIR*, pp. 119-126.

[30] Jeong, K.S., Myaeng, S.H., Lee, J.S, and Choi, K.S., (1999) Automatic identification and back-transliteration of foreign words for

information retrieval, *Information Processing and Management,* 35(4), pp. 523-540.

[31] Ji, Heng (2009) Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks, *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009*, pages34–37.

[32] Koehn, P., Och, F.J., Marcus, D., (2003) Statistical phrase-based translation, In *Proceedings of HLT-NAACL*, pp. 48-54.

[33] Koehn, P. (2009) Statistical Machine Translation. Cambridge University Press.

[34] Kraaij, W., Nie, J.Y., and Simard, M. (2003). Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*, 29(3): 381-420.

[35] Kumano, T. and Tanaka, H., Tokunaga, T. (2007) Extracting phrasal alignments from comparable corpora by using joint probability SMT model. 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07).

[36] Kuo, J.S., Li, H., and Yang Y.K (2006). Learning Transliteration Lexicon from the Web. In *the Proceedings of COLING/ACL*, pp.1129-1136

[37] Lam, W., Chan, S.K., and Huang, R. (2007). Named Entity Translation Matching and Learning: With Application for Mining Unseen Translations. *ACM Transactions on Information Systems*, 25(1), pp.

[38] Liu, Y., Jin R. and Chai, Joyce Y. (2005). A maximum coherence model for dictionary-based cross-language information retrieval, In *Proceedings of SIGIR conf.*, pp. 536-543.

[39] Lu, W. Chien, L.F. and Lee, H. (2004). Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems,* Vol.22, pp. 242-269.

[40] Ma, X. and Liberman, M., (1999). Bits: A Method for Bilingual Text Search over the Web. *Proceedings of Machine Translation Summit VII.*

[41] Munteanu, D. S., Marcu, D. (2005) Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. 2005. *Computational Linguistics*. 31(4). pp: 477-504.

[42] Munteanu, D. S. and Marcu D. (2006). Extracting parallel subsentential fragments from non-parallel corpora. *ACL*, pp. 81–88.

[43] Nagata, M., Saito, T., and Suzuki, K. (2001). Using the web as a bilingual dictionary. In *Proceedings of the Workshop on Data-Driven Methods in Machine Translation* (with ACL Conf.), pp. 1-8.

[44] Nie, J.Y., Cai, J. (2001) Filtering parallel corpora of web pages, IEEE symposium on NLP and Knowledge Engineering, pp. 453-458.

[45] Nie, J.Y., Simard, M., Isabelle, P., Durand, R. (1999) Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web, In *Proceedings of SIGIR Conf.*, pp. 74-81

[46] Och, F., and Ney, H. (2002) Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *ACL*, pp. 295-302

[47] Och, F. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL*. pp. 160-67

[48] Oumohmed, A.I., Mignotte, M., Nie, J.Y. (2005) Semantic-Based Cross-Media Image Retrieval, *Pattern Recognition and Image Analysis: Third International Conference on Advances in Pattern Recognition (ICAPR)*, LNCS 3687, pp. 414-423.

[49] Potthast, M., Stein, B., Anderka, M. (2008) A Wikipedia-based Multilingual Retrieval Model. *ECIR,* LNCS 4956, pp. 522–530.

[50] Qu, Y., Grefenstette, G., and Evans, D. A. (2003). Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of SIGIR Conference*, pp. 353-360.

[51] Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

[52] Resnik, P., (1999) Mining the Web for Bilingual Text, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99).

[53] Resnik P. and Smith. N.A. (2003) The Web as a Parallel Corpus, *Computational Linguistics,* 29(3), pp. 349-380, September 2003.

[54] Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of SIGIR Conf.,* pp. 58-65.

[55] Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K. (2007) Performing cross-language retrieval with Wikipedia, CLEF-2007 (http://www.clef-campaign.org/2007/working_notes/schonhofenCLEF2007.pdf)

[56] Shi, L., Niu, C., Zhou, M., and Gao, J. (2006) A DOM Tree Alignment Model for Mining Parallel Data from the Web, *ACL,* pp. 489-496.

[57] Smith, J. R., Quirk, C., and Toutanova, K. (2010) Extracting parallel sentences from comparable corpora using document level alignment. *HLT*, pp. 403–411

[58] Sproat, R., Tao, T., Zhai, C. (2006) Named Entity Transliteration with Comparable Corpora. In *Proceedings of ACL.*

[59] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola (2006) A study on automatic creation of a comparable document collection in cross-language information retrieval, Journal of Documentation, Vol. 62 No. 3, pp. 372-387

[60] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.* 25, 1, Article 4.

[61] Utiyama M. and Isahara, H. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. *ACL*, pp. 72–79.

[62] Jinxi Xu, W. Bruce Croft (1996) Query Expansion Using Local and Global Document Analysis. *SIGIR*, pp. 4-11

[63] Yang, Christopher C., and Kar Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *Journal of the American Society for Information Science and Technology*, 54(8), pp. 730–742.

[64] Zhang, Y. and Vines, P. (2004). Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proceedings of SIGIR Conf.*, pp.162-169.

[65] Zhang, Y., Huang, F., Vogel, S. (2005) Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion, *SIGIR*, pp. 669-670.

[66] Zhao, B., and Vogel, S. (2002). Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In Proceedings of IEEE international conference on data mining, pages 745-750.

Chapter 11

TEXT MINING IN MULTIMEDIA

Zheng-Jun Zha
School of Computing, National University of Singapore
zhazj@comp.nus.edu.sg

Meng Wang
School of Computing, National University of Singapore
wangm@comp.nus.edu.sg

Jialie Shen
Singapore Management University
jlshen@smu.edu.sg

Tat-Seng Chua
School of Computing, National University of Singapore
chuats@comp.nus.edu.sg

Abstract A large amount of multimedia data (e.g., image and video) is now available on the Web. A multimedia entity does not appear in isolation, but is accompanied by various forms of metadata, such as surrounding text, user tags, ratings, and comments etc. Mining these textual metadata has been found to be effective in facilitating multimedia information processing and management. A wealth of research efforts has been dedicated to text mining in multimedia. This chapter provides a comprehensive survey of recent research efforts. Specifically, the survey focuses on four aspects: (a) surrounding text mining; (b) tag mining; (c) joint text and visual content mining; and (d) cross text and visual content mining. Furthermore, open research issues are identified based on the current research efforts.

Keywords: Text Mining, Multimedia, Surrounding Text, Tagging, Social Network

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_11,

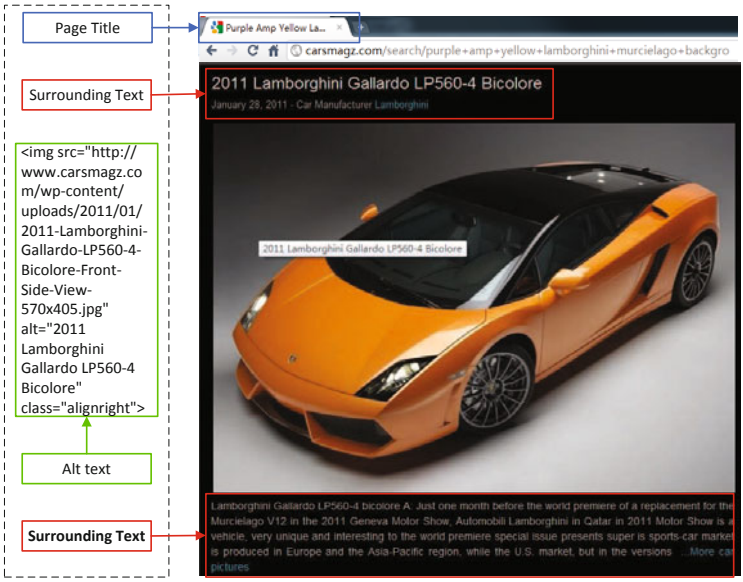

Figure 11.1. Illustration of textual metadata of an embedded image in a Web page.

1. Introduction

Lower cost hardware and growing communications infrastructure (e.g. Web, cell Phones, etc.) have led to an explosion in the availability of ubiquitous devices to produce, store, view and exchange multimedia entities (images, videos). A large amount of image and video data are now available. Take one of the most popular photo sharing services Flickr [1] as example, it has accumulated several billions of images. Another example is Youtube [2], which is a video sharing Web site that is hosting billions of videos. As the largest photo sharing site, Facebook [3] currently stores hundreds of hundreds of billions of photos.

On the other hand, a multimedia entity does not appear in isolation but is accompanied by various forms of textual metadata. One of the most typical examples is the surrounding text appearing around the embedded images or videos in the Web page (See Figure 11.1). With recent proliferation of social media sharing services, the newly emerging textual meatadata include user tags, ratings, comments, as well as

[1] http://www.flickr.com/
[2] http://www.youtube.com/
[3] http://www.facebook.com/

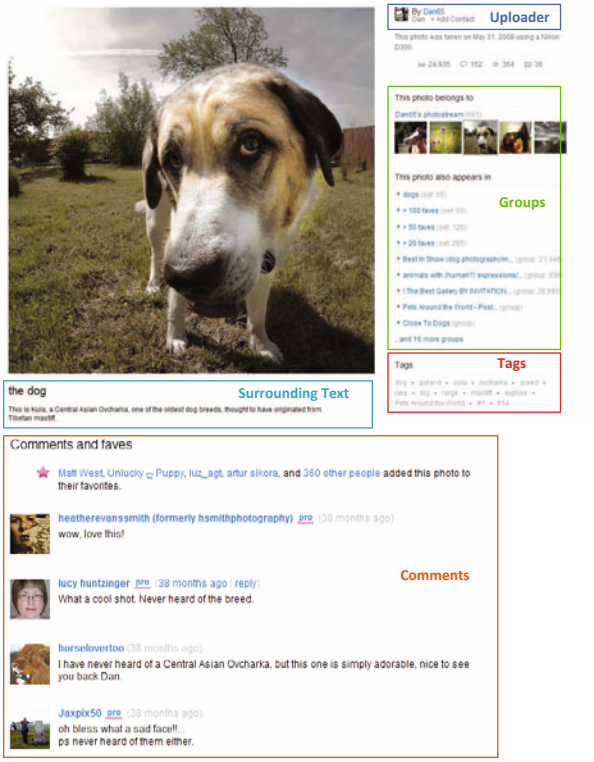

Figure 11.2. Illustration of textual metadata of an image on a photo sharing Web site.

the information about the uploaders and their social network (See Figure 11.2). These metadata, in particular the tags, have been found to be an important resource for facilitating multimedia information processing and management. Given the wealth of research efforts that has been done, there have been various studies in multimedia community on the mining of textual metadata. In this chapter, a multimedia entity refers to an image or a video. For the sake of simplicity and without lost of generality, we use the term image to refer to multimedia entity for the rest of this chapter.

In this chapter, we first review the related works on mining surrounding text for image retrieval as well as the recent research efforts that explore surrounding text for image annotation and clustering in Section 2. In Section 3, we provide a literature review on tag mining and show that the main focus of existing tag mining works includes three aspects: tag ranking, tag refinement, and tag information enrichment. In

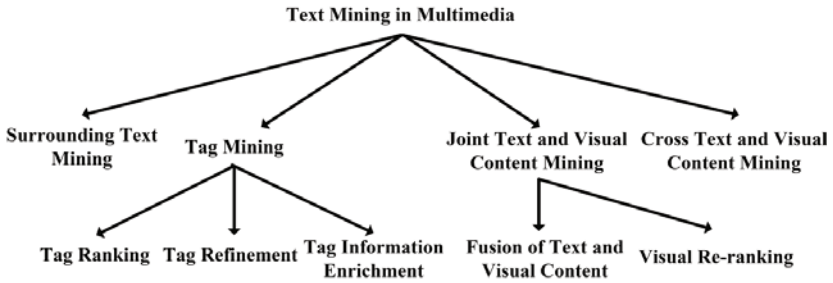

Figure 11.3. A taxonomy consisting of the research works reviewed in this chapter.

Section 4, we survey the recent progress in integrating textual metadata and visual content. We categorize the exiting works into two categories: the fusion of text and visual content as well as visual re-ranking. In Section 5, we provide a detailed discussion on recent research on cross text and visual content mining. We organize all the works reviewed in this chapter into a taxonomy as shown in Figure 11.3. The taxonomy provides an overview of state-of-the-art research and helps us to identify open research issues to be presented in Section 6.

2. Surrounding Text Mining

In order to enhance the content quality and improve user experience, many hosting Web pages include different kinds of multimedia entities, like image or video. These multimedia entities are frequently embedded as part of the text descriptions which we called the surrounding text. While there is no standard definition, surrounding text generally refers to the text consisting of words, phrases or sentences that surrounds or close to the embedded images, such as those that appear at the top, below, left or right region of images or connected via Web links. The effective use of surrounding texts is becoming increasingly important for multimedia retrieval. However, developing effective extraction algorithm for the comprehensive analysis of surrounding text has been a very challenging task. In many cases, automatically determining which page region is more relevant to the image than the others could be difficult. Moreover, how large the region nearby should be considered is still an open question. Further, the quality of surrounding texts could be low and inconsistent. These problems make it very hard to directly apply the surrounding text information to facilitate accurate retrieval. Thus, refinement process or combining it with other cues is essential.

The earliest efforts on modeling and analyzing surrounding texts to facilitate multimedia retrieval occurred in the 1990s. AltaVista's A/V

Photo Finder applies textual and visual cues to index image collections [1]. The indexing terms are precomputed based on the HTML documents containing the Web images. With a similar approach, the WebSeer system harvests the information for indexing Web images from two different sources: the related HTML text and the embedded image itself [12]. It extracts keywords from page title, file name, caption, alternative text, image hyperlinks, and body text titles. A weight is calculated for each keyword based on its location inside a page. In PICITION system [40], an interesting approach is developed to exploit both textual and visual information to index a pictorial database. Image captions are used as an important cue to identify faces appearing in a related newspaper photograph. The empirical study based on a data set containing 50 pictures and captions obtained from the $Buffalo\ News$ and the $New\ York\ Times$ is used to demonstrate the effectiveness of the PICITION system. While the system can be successfully adopted for accessing photographs in newspaper or magazine, it is not straightforward to apply it for Web image retrieval.

In [39], Smith and Chang proposed the WebSeek framework designed to search images from the Web. The key idea is to analyze and classify the Web multimedia objects into a predefined taxonomy of categories. Thus, an initial search can be performed to explore a catalog associated with the query terms. The image attribute (e.g., color histogram for images) is then computed for similarity matching within the category.

Besides its efficacy in image retrieval, surrounding text has been explored for image annotation recently. Feng et al. presented a bootstrapping framework to label and search Web images based on a set of predefined semantic concepts [9]. To achieve better annotation effectiveness, a co-training scheme is designed to explore the association between the text features computed using corresponding HTML documents and visual features extracted from image content. Observing that the links between the visual content and the surrounding texts can be modeled via Web page analysis, a novel method called Iterative Similarity Propagation is proposed to refine the closeness between the Web images and their annotations [50]. On the other hand, it is not hard to find that images from the same cluster may share many similar characteristics or patterns with respect to relevance to information needs. Consequently, accurate clustering is a very crucial technique to facilitate Web multimedia search and many algorithms have recently been proposed based on the analysis of surrounding texts and low level visual features [3][13][34]. For example, Cai et al. [3] proposed a hierarchical clustering method that exploits visual, textual, and link analysis. A webpage is partitioned into blocks, and the textual and link information

of an image are extracted from the block containing that image. By using block-level link analysis techniques, an image graph is constructed. They then applied spectral techniques to find a Euclidean embedding of the images. As a result, each image has three types of representations: visual feature, textual feature, and graph-based representation. Spectral clustering techniques are employed to cluster search results into various clusters. Gao et al. [13] and Rege et al. [34] used a tripartite graph to model the relations among visual features, images and their surrounding text. The clustering is performed by partitioning this tripartite graph.

3. Tag Mining

In newly emerging social media sharing services, such as the Flickr and Youtube, users are encouraged to share multimedia data on the Web and annotate content with tags. Here a tag is referred to as a descriptive keyword that describes the multimedia content at semantic or syntactic level. These tags have been found to be an important resource for multimedia management and have triggered many innovative research topics [61][51][38][36]. For example, with accurate tags, the retrieval of multimedia content can be easily accomplished. The tags can be used to index multimedia data and support efficient tag-based search. Nowadays, many online media repositories, such as Flickr and Youtube, support tag-based multimedia search. However, since the tags are provided by grassroots Internet users, they are often noisy and incomplete and there is still a gap between these tags and the actual content of the images[20][26][48]. This deficiency has limited the effectiveness of tag-based applications.

Recently, a wealth of research has been proposed to enhance the quality of human-provided tags. The existing works mainly focus on the following three aspects: (a) tag ranking, which aims to differentiate the tags associated with the images with various levels of relevance; (b) tag refinement with the purpose to refine the unreliable human-provided tags; and (c) tag information enrichment, which aims to supplement tags with additional information [26]. In this section, we present a comprehensive review of existing tag ranking, tag refinement, and tag information enrichment methods.

3.1 Tag Ranking

As shown in [25], the relevance level of the tags cannot be distinguished from the tag list of an image. The lack of relevance information in the tag list has limited the application of tags. Recently, tag ranking has been studied to infer the relevance levels of tags associated with an

Figure 11.4. Examples of of tag refinement. The left side of the figure shows the original tags while the right side shows the refined tags. The technique is able to remove irrelevant tags and add relevant tags to obtain better description of multimedia contents.

image. As a pioneering work, Liu et al. [25] proposed to estimate tag relevance scores using kernel density estimation, and then employ random walk to boost this primary estimation. Li et al. [22] proposed a data driven method for tag ranking. They learned the relevance scores of tags by a neighborhood voting approach. Given an image and one of its associated tag, the relevance score is learned by accumulating the votes from the visual neighbors of the image. They then extended the work to multiple visual spaces [23]. They learned the relevance scores of tags and ranked them by neighborhood voting in different feature spaces, and the results are aggregated with a score fusion or rank fusion method. Different aggregation methods have been investigated, such as the average score fusion, Borda count and RankBoost. The results show that a simple average fusion of scores is already able to perform closed to supervised fusion methods like RankBoost.

3.2 Tag Refinement

User-provided tags are often noisy and incomplete. The study in [20] shows that when a tag appears in a Flickr image, there is only about a 50% chance that the tag is really relevant, and the study in [38] shows that more than half of Flickr images are associated with less than three tags. Tag refinement technologies are proposed aiming at obtaining more

accurate and complete tags for multimedia description, as shown in Figure 11.4.

A lot of tag refinement approaches have been developed based on various statistical learning techniques. Most of them are based on the following three assumptions.

- The refined tags should not change too much from those provided by the users. This assumption is usually used to regularize the tag refinement.
- The tags of visually similar images should be closely related. This is a natural assumption that most automatic tagging methods are also built upon.
- Semantically close or correlative tags should appear with high correlation. For example, when a tag "sea" exists for an image, the tags "beach" and "water" should be assigned with higher confidence while the tag "street" should have low confidence.

For example, Chen et al. [6] first trained a SVM classifier for each tag with the loosely labeled positive and negative samples. The classifiers are used to estimate the initial relevance scores of tags. They then refined the scores with a graph-based method that simultaneously considers the similarity between images and semantic correlation among tags. Xu et al. [52] proposed a tag refinement algorithm from topic modeling point of view. A new graphical model named regularized latent Dirichlet allocation (rLDA) is presented to jointly model the tag similarity and tag relevance. Zhu et al. [64] proposed a matrix decomposition method. They used a matrix to represent the image-tag relationship: the (i, j)-th element is 1 if the i-th image is associated with the j-th tag, and 0 otherwise. The matrix is then decomposed into a refined matrix plus an error matrix. They enforced the error matrix to be sparse and the refined matrix to follow three principles: (a) let the matrix be low-rank; (b) if two images are visually similar, the corresponding rows are with high correlation; and (c) if two tags are semantically close, the corresponding vectors are with high correlation. Fan et al. [8] grouped images with a target tag into clusters. Each cluster is regarded as a unit. The initial relevance scores of the clusters are estimated and then refined by a random walk process. Liu et al. [24] adopted a three-step approach. The first step filters out tags that are intrinsically content-unrelated based on the ontology in WordNet. The second step refines the tags based on the consistency of visual similarity and semantic similarity of images. The last step performs tag enrichment, which expands the tags with their appropriate synonyms and hypericum.

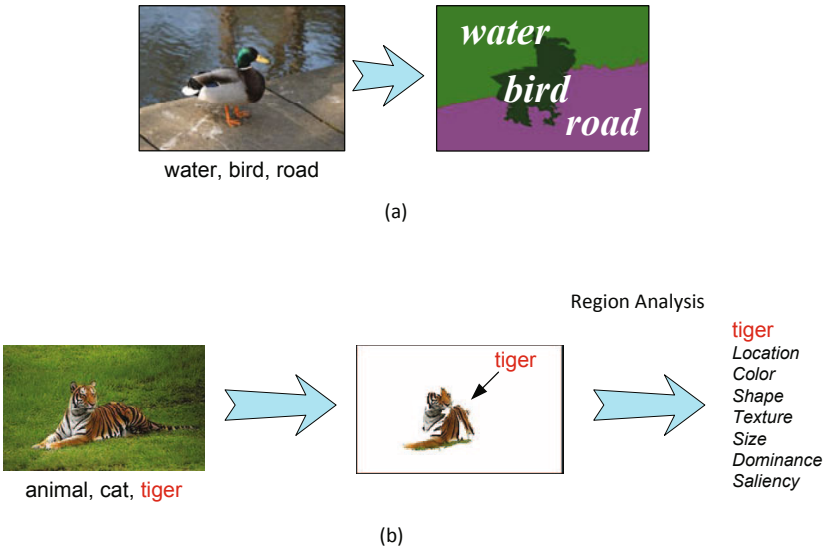

Figure 11.5. (a) An example of tag localization, which finds the regions that the tags describe. (b) An illustration of tag information enrichment. It first finds the corresponding region of the target tag and then analyze the properties of the region.

3.3 Tag Information Enrichment

In the manual tagging process, generally human labelers will only assign appropriate tags to multimedia entities without any additional information, such as the image regions depicted by the corresponding tags. But by employing computer vision and machine learning technologies, certain information of the tags, such as the descriptive regions and saliency, can be automatically obtained. We refer to these as tag information enrichment.

Most existing works employ the following two steps for tag information enrichment. First, tags are localized into regions of images or sub-clips of videos. Second, the characteristics of the regions or sub-clips are analyzed, and the information about the tags is enriched accordingly. Figure 11.5 (a) illustrates the examples of tag localization for image and video data. Liu et al. [28] proposed a method to locate image tags to corresponding regions. They first performed over-segmentation to decompose each image into patches and then discovered the relationship between patches and tags via sparse coding. The over-segmented regions are then merged to accomplish the tag-to-region process. Liu et al. extended the approach based on image search [29]. For a tag of the target image, they collected a set of images by using the tag as query

with an image search engine. They then learned the relationship between the tag and the patches in this image set. The selected patches are used to reconstruct each candidate region, and the candidate regions are ranked based on the reconstruction error. Liu et al. [27] accomplished the tag-to-region task by regarding an image as a bag of regions and then performed tag propagation on a graph, in which vertices are images and edges are constructed based on the visual link of regions. Feng et al. [10] proposed a tag saliency learning scheme, which is able to rank tags according to their saliency levels to an image's content. They first located tags to images' regions with a multi-instance learning approach. In multi-instance learning, an image is regarded as a bag of multiple instances, i.e., regions [58]. They then analyzed the saliency values of these regions. It can provide more comprehensive information when an image is relevant to multiple tags, such as those describing different objects in the image. Yang et al. [55] proposed a method to associate a tag with a set of properties, including location, color, texture, shape, size and dominance. They employed a multi-instance learning method to establish the region that each tag is corresponding to, and the region is then analyzed to establish the properties, as shown in Figure 11.5 (b). Sun and Bhowmick [41] defined a tag's visual representativeness based on a large image set and the subset that is associated with the tag. They employed two distance metrics, cohesion and separation, to estimate the visual representativeness measure.

Ulges et al. [43] proposed an approach to localize video-level tags to keyframes. Given a tag, it regards whether a keyframe is relevant as a latent random variable. An EM-style process is then adopted to estimate the variables. Li et al. [21] employed a multi-instance learning approach to accomplish the video tag localization, in which video and shot are regarded as bag and shot, respectively.

By supplementing tags with additional information, a lot of tag-based applications can be facilitated, such as tag-based image/video retrieval and intelligent video browsing etc.

4. Joint Text and Visual Content Mining

Beyond mining pure textual metadata, researchers in multimedia community have started making progress in integrating text and content for multimedia retrieval via joint text and content mining. The integration of text and visual content has been found to be more effective than exploiting purely text or visual content separately. The joint text and content mining in multimedia retrieval often comes down to finding effective mechanisms for fusing multi-modality information from textual

metadata and visual content. Existing research efforts can generally be categorized into four paradigms: (a) linear fusion; (b) latent-space-based fusion; (c) graph-based fusion; and (d) visual re-ranking that exploits visual information to refine text-based retrieval results. In this section, we first briefly review linear, latent space based, and graph based fusion methods and then provide comprehensive literature review on visual re-ranking technology.

Linear fusion combines the retrieval results from various modalities linearly [18][4][31]. In [18], visual content and text are combined in both online learning stage with relevance feedback and offline keyword propagation. In [31], linear, max, and average fusion strategies are employed to aggregate the search results from visual and textual modalities. Chang et al. [4] adopted a query-class-dependent fusion approach. The critical task in linear fusion is the estimation of fusion weights of different modalities. A certain amount of training data is usually required for estimating these weights. The latent space based fusion assumes that there is a latent space shared by different modalities and thus unify different modalities by transferring the features of these modalities into the shared latent space [63][62]. For example, Zhao et al. [63] adopted the Latent Semantic Indexing (LSI) method to fuse text and visual content. Zhang et al. [62] proposed a probabilistic context model to explicitly exploit the synergy between text and visual content. The synergy is represented as a hidden layer between the image and text modalities. This hidden layer constitutes the semantic concepts to be annotated through a probabilistic framework. An Expectation-Maximization (EM) based iterative learning procedure is developed to determine the conditional probabilities of the visual features and the words given a hidden concept class. Latent space based methods usually require a large amount of training samples for learning the feature mapping from each modality into the unified latent space. Graph based approach [49] first builds the relations between different modalities, such as relations between images and text using the Web page structure. The relations are then utilized to iteratively update the similarity graphs computed from different modalities. The difficulty of creating similarity graphs for billions of images on the Web makes this approach insufficiently scalable.

4.1 Visual Re-ranking

Visual re-ranking is emerging as one of the promising technique for automated boosting of retrieval precision [42] [30] [55]. The basic functionality is to reorder the retrieved multimedia entities to achieve the optimal rank list by exploiting visual content in a second step. In par-

ticular, given a textual query, an initial list of multimedia entities is returned using the text-based retrieval scheme. Subsequently, the most relevant results are moved to the top of the result list while the less relevant ones are reordered to the lower ranks. As such, the overall search precision at the top ranks can be enhanced dramatically. According to the statistical analysis model used, the existing re-ranking approaches can roughly be categorized into three categories including the clustering based, classification based and graph based methods.

Cluster analysis is very useful to estimate the inter-entity similarity. The clustering based re-ranking methods stem from the key observation that a lot of visual characteristics can be shared by relevant images or video clips. With intelligent clustering algorithms (e.g., mean-shift, K-means, and K-medoids), initial search results from text-based retrieval can be grouped by visual closeness. One good example of clustering based re-ranking algorithms is an Information Bottle based scheme developed by Hsu et al. [16]. Its main objective is to identify optimal clusters of images that can minimize the loss of mutual information. The cluster number is manually configurated to ensure the each cluster contains the same number of multimedia entities (about 25). This method was evaluated using the TRECVID 2003-2005 data and significant improvements were observed in terms of MAP measures. In [19], a fast and accurate scheme is proposed for grouping Web image search results into semantic clusters. For a given query, a few related semantic clusters are identified in the first step. Then, the cluster names relating to query are derived and used as text keywords for querying image search engine. The empirical results from a set of user studies demonstrate an improvement in performance over Google image search results. It is not hard to show that the clustering based re-ranking methods can work well when the initial search results contain many near-duplicate media documents. However, for queries that return highly diverse results or without clear visual patterns, the performance of the clustering-based methods is not guaranteed. Furthermore, the number of clusters has large impact on the final effectiveness of the algorithms. However, determining the optimal cluster number automatically is still an open research problem.

In the classification based methods, visual re-ranking is formulated as a binary classification problem aiming to identify whether each search result is relevant or not. The major process for result list reordering consists of three major steps: (a) the selection of pseudo-positive and pseudo-negative samples; (b) use the samples obtained in step (a) to train a classification scheme; and (c) reorder the samples according to their relevance scores given by the trained classifier. For existing classification methods, pseudo relevance feedback (PRF) is applied to select the

training examples. It assumes that: (a) a limited number of top-ranked entities in the initial retrieval results are highly relevant to the search queries; and (b) automatic local analysis over the entities can be very helpful to refine query representation. In [54], the query images or video clip examples are used as the pseudo-positive samples. The pseudo-negative samples are selected from either the least relevant samples in the initial result list or the databases that contain less samples related to the query. The second step of the classification based methods aim to train classifiers and a wide range of statistical classifiers can be adopted. They include the Support Vector Machine (SVM) [54], Boosting [53] and ListNet [57]. The main weakness for the classification based methods is that the number and quality of training data required play a very important role in constructing effective classifiers. However, in many real scenarios, the training examples obtained via PRF are very noisy and might not be adequate for training effective classifier. To address this issue, Fergus et al. [11] used RANSAC to sample a training subset with a high percentage of relevant images. A generative constellation model is learned for the query category while a background model is learned from the query "things". Images are re-ranked based on their likelihood ratio. Observing that discriminative learning can lead to superior results, Schroff et al. [35] first learned a query independent text based re-ranker. The top ranked results from the text based re-ranking are then selected as positive training examples. Negative training examples are picked randomly from the other queries. A binary SVM classifier is then used to re-rank the results on the basis of visual features. This classifier is found to be robust to label noise in the positive training set as long as the non-relevant images are not visually consistent. Better training data can be obtained from online knowledge resources if the set of queries restricted. For instance, Wang et al. [44] learned a generative text model from the query's Wikipedia [4] page and a discriminative image model from the Caltech [15] and Flickr data sets. Search results are then re-ranked on the basis of these learned probability models. Some user interactions are required to disambiguate the query.

Graphs provide a natural and comprehensive way to explore complex relations between data at different levels and have been applied to a wide range of applications [59][46][47][60]. With the graph based re-ranking methods, the multimedia entities in top ranks and their associations/dependencies can be represented as a collection of nodes (vertices) and edges. The local patterns or salient features discover using graph

[4] http://www.wikipedia.org/

analysis are very helpful to improve effectiveness of rank lists. In [16], Hsu et al. modeled the re-ranking process as a random walk over the context graph. In order to effectively leverage the retrieved results from text search, each sample corresponds to a "dongle" node containing ranking score based on text. For the framework, edges between "dongle" nodes are weighted with multi-modal similarities. In many cases, the structure of large scale graphs can be very complex and this easily makes related analysis process very expensive in terms of computational cost. Thus, Jing and Baluja proposed a VisualRank framework to efficiently model similarity of Google image search results with graph [17]. The framework casts the re-ranking problem as random walk on an affinity graph and reorders images according to the visual similarities. The final result list is generated via sorting the images based on graph nodes' weights. In [42], Tian et al., presented a Bayesian video search re-ranking framework formulating the re-ranking process as an energy minimization problem. The main design goal is to optimize the consistency of ranking scores over visually similar videos and minimize the disagreement between the optimal list and the initial list. The method achieves a consistently better performance over several earlier proposed schemes on the TRECVID 2006 and 2007 data sets. The graph based re-ranking algorithms mentioned above generally do not consider any initial supervision information. Thus, the performance is significantly dependent on the statistical properties of top ranked search results. Motivated by this observation, Wang et al, proposed a semi-supervised framework to refine the text based image retrieval results via leveraging the data distribution and the partial supervision information obtained from the top ranked images [45]. Indeed, graph analysis has been shown to be a very powerful tool for analyzing and identifying salient structure and useful patterns inside the visual search results. With recent progresses in graph mining, this research stream is expected to continue to make important contributions to improve visual re-ranking from different perspectives.

5. Cross Text and Visual Content Mining

Although the joint text and visual content mining approaches described above facilitate image retrieval, they require that the test images have associated text modality. However, in some real world applications, images may not always have associated text. For example, most surveillance images/videos in in-house repository are not accompanied with any text. Even on social media Website such as the Flickr, there exist a substantial number of images without any tags. In such cases, joint

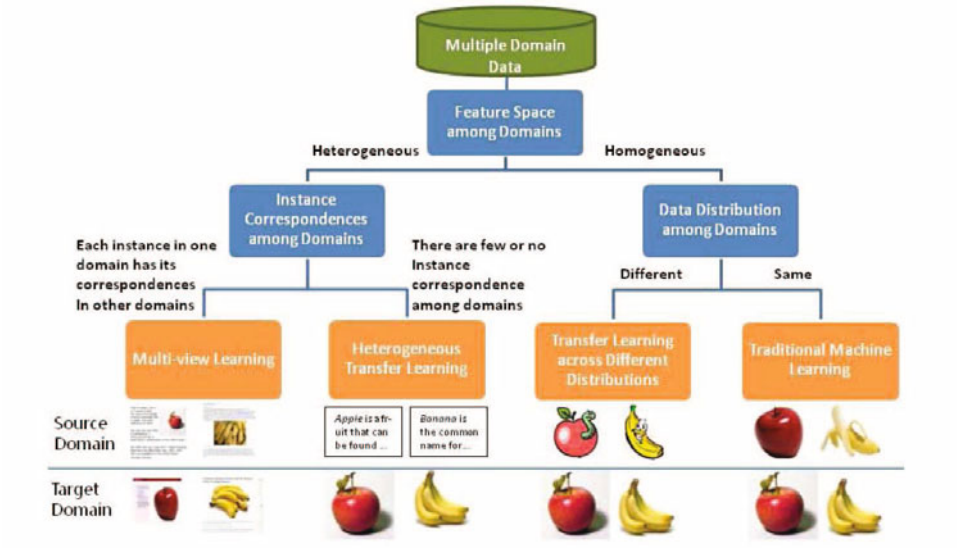

Figure 11.6. An illustration of different types of learning paradigms using image classification/clustering in the domains of apple and banana. Adapted from [56].

text and visual content mining cannot be applied due to missing text modality.

Recently, cross text and visual content mining has been studied in the context of transfer learning techniques. This class of techniques emphasizes the transferring of knowledge across different domains or tasks [32]. Cross text and visual content mining does not require that a test image has an associated text modality, and is thus beneficial to dealing with the images without any text by propagating the semantic knowledge from text to images [5]. It is also motivated by two observations. First, visual content of images is much more complicated than the text feature. While the textual words are easier to interpret, there exist a tremendous semantic gap between visual content and high-level semantics. Second, image understanding becomes particularly challenging when only a few labeled images are available for training. This is a common challenge, since it is expensive and time-consuming to obtain labeled images. On the contrary, labeled/unlabeled text data are relatively easier to collect. For example, millions of categorized text articles are freely available in Web

[5]Cross text and visual content can also facilitate text understanding in special cases by propagating knowledge from images to text.

text collections, such as Wikipedia, covering a wide range of topics from culture and arts, geography and places, history and events, to natural and physical science. A large number of Wikipedia articles are indexed by thousands of categories in these topics [33]. This provides abundant labeled text data. Thus, it is desirable to propagate semantic knowledge from text to images to facilitate image understanding. However, it is not trivial to transfer knowledge between various domains/tasks due to the following challenges:

- The target data may be drawn from a distribution different from the source data.
- The target and source data may be in different feature spaces (e.g., image and text) and there may be no correspondence between instances in these spaces.
- The target and source tasks may have different output spaces.

While the traditional transfer learning techniques focus on the distribution variance problem, the recent proposed heterogenous transfer learning approaches aim to tackle both the distribution variance and heterogenous feature space problems [56][7][65][33], or all the three challenges listed above [37]. Figure 11.6 from [56] presents an intuitive illustration of four learning paradigms, including traditional machine learning, transfer learning across different distributions, multi-view learning and heterogenous transfer learning. As we can see, heterogenous transfer learning is usually much more challenging due to the unknown correspondence across the distinct feature spaces. In order to learn the underlying correspondence for knowledge transformation, a "semantic bridge" is required. The "semantic bridge" can be obtained from the co-occurrence information between text and images or the linkage information in social media networks. For example, while the traditional webpages provide the co-occurrence information between text and images, the social media sites contain a large number of linked information between different types of entities, such as the text articles, tags, posts, images and videos. This linkage information provide a "semantic bridge" to learn the underlying correspondence [2].

Most existing works exploit the tag information that provide text-to-image linking information. As a pioneering work, Dai et al. [7] showed that such information can be effectively leveraged for transferring knowledge between text and images. The key idea of [7] is to construct a correspondence between the images and the auxiliary text data with the use of tags. Probabilistic latent semantic analysis (PLSA) model is employed to construct a latent semantic space which can be used for

transferring knowledge. Chen et al. [56] proposed the concept of heterogeneous transfer learning and applied it to improve image clustering by leveraging auxiliary text data. They collected annotated images from the social web, and used them to construct a text to image mapping. The algorithm is referred to as aPLSA (Annotated Probabilistic Latent Semantic Analysis). The key idea is to unify two different kinds of latent semantic analysis in order to create a bridge between the text and images. The first kind of technique performs PLSA analysis on the target images, which are converted to an image instance-to-feature co-occurrence matrix. The second kind of PLSA is applied to the annotated image data from social Web, which is converted into a text-to-image feature co-occurrence matrix. In order to unify those two separate PLSA models, these two steps are done simultaneously with common latent variables used as a bridge linking them. It has been shown in [5] that such a bridging approach leads to much better clustering results. Zhu et al. [65] discussed how to create the connections between images and text with the use of tag data. They showed how such links can be used more effectively for image classification. An advantage of [65] is that it exploits unlabeled text data instead of labeled text as in [7].

In contrast to these methods that exploit tag information to link images and auxiliary text articles, Qi et al. [33] proposed to learn a "translator" which can directly establish the semantic correspondence between text and images even if they are new instances of the image data with unknown correspondence to the text articles. This capability increase the flexibility of the approach and makes it more widely applicable. Specifically, they created a new topic space into which both the text and images are mapped. A translator is then learned to link the instances across heterogeneous text and image spaces. With the resultant translator, the semantic labels can be propagated from any labeled text corpus to any new image by a process of cross-domain label propagation. They showed that the learned translator can effectively convert the semantics from text to images.

6. Summary and Open Issues

In this chapter, we have reviewed the active research on text mining in multimedia community, including surrounding text mining, tag mining, joint text and visual content mining, and cross text and visual content mining. Although research efforts in this filed have made great progress in various aspects, there are still many open research issues that need to be explored. Some examples are listed and discussed as follows.

Joint text and visual content multimedia ranking

Despite the success of visual re-ranking in multimedia retrieval, visual re-ranking only employs the visual content to refine text-based retrieval results; visual content has not been used to assist in learning the ranking model of search engine, and sometimes it is only able to bring in limited performance improvements. In particular, if text-based ranking model is biased or over-fitted, re-ranking step will suffer from the error that is propagated from the initial results, and thus the performance improvement will be negatively impacted. Therefore, it is worthwhile to simultaneously exploit textual metadata and visual content to learn a unified ranking model. A preliminary work has been done in [14], where a content-aware ranking model is developed to incorporate visual content into text-based ranking model learning. It shows that the incorporation of visual content into ranking model learning can result in a more robust and accurate ranking model since noise in textual features can be suppressed by visual information.

Scalable text mining for large-scale multimedia management

Despite of the success of existing text mining in multimedia, most existing techniques suffer from difficulties in handling large-scale multimedia data. Huge amount of training data or high computation powers are usually required by existing methods to achieve acceptable performance. However, it is too difficult, or even impossible, to meet this requirement in real-world applications. Thus there is a compelling need to develop scalable text mining techniques to facilitate large-scale multimedia management.

Multimedia social network mining

In recent years, we have witnessed the emergence of multimedia social network communities like Napster [6], Facebook [7], and Youtube, where millions of users and billions of multimedia entities form a large-scale multimedia social network. Multimedia social networking is becoming an important part of media consumption for Internet users. It brings in new and rich metadata, such as user preferences, interests, behaviors, social relationships, and social network structure etc. These information present new potential for advancing current multimedia analysis

[6] http://music.napster.com/
[7] http://www.facebook.com/

techniques and also trigger diverse multimedia applications. Numerous research topics can be explored, including (a) the combination of conventional techniques with information derived from social network communities; (b) fusion analysis of content, text, and social network data; and (c) personalized multimedia analysis in social networking environments.

Acknowledgements

This work was in part supported by A*Star Research Grant R-252-000-437-305 and NRF (National Research Foundation of Singapore) Research Grant R-252-300-001-490 under the NExT Center.

References

[1] Altavista's a/v photo finder. http://www.altavista.com/sites/search/simage.

[2] C. C. Aggarwal, H. Wang. *Text Mining in Social Networks*. Social Network Data Analytics, Springer, 2011.

[3] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the ACM Conference on Multimedia*, 2004.

[4] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia university trecvid-2006 video search and high-level feature extraction. In *Proceedings of NIST TRECVID workshop*, 2006.

[5] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proceedings of the ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[6] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[7] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across difference feature spaces. In *NIPS*, pages 353–360, 2008.

[8] J. Fan, Y. Shen, N. Zhou, and Y. Gao. Harvesting large-scaleweakly-tagged image databases from the web. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[9] H. Feng, R. Shi, and T.-S. Chua. A bootstrapping framework for annotating and retrieving www images. In *Proceedings of the ACM Conference on Multimedia*, 2004.

[10] S. Feng, C. Lang, and D. Xu. Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking. In *Proceedings of International Conference on Image and Video Retrieval*, 2010.

[11] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proceedings of the European Conference on Computer Vision*, 2004.

[12] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. Technical report, University of Chicago, Computer Science Department, 1996.

[13] B. Gao, T.-Y. Liu, Q. Tao, X. Zheng, Q. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the ACM Conference on Multimedia*, 2005.

[14] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Content-aware ranking for visual search. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2010.

[15] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[16] W. Hsu, L. Kennedy, , and S.-F. Chang. Reranking methods for visual search. *IEEE Multimedia*, 14:14–22, 2007.

[17] F. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1877–1890, 2008.

[18] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 2005.

[19] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: Web image search results clustering. In *Proceedings of the ACM Conference on Multimedia*, pages 377–384, 2006.

[20] L. S. Kennedy, S. F. Chang, and I. V. Kozintsev. To search or to label? predicting the performance of search-based automatic image classifiers. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2006.

[21] G. Li, M. Wang, Y. T. Zheng, Z.-J. Zha, H. Li, and T.-S. Chua. Shottagger: Tag location for internet videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.

[22] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *Pattern Recognition Letters*, 11(7), 2009.

[23] X. Li, C. G. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, 2010.

[24] D. Liu, X. C. Hua, M. Wang, and H. Zhang. Image retagging. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[25] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the International Conference on World Wide Web*, 2009.

[26] D. Liu, X.-S. Hua, and H.-J. Zhang. Content-based tag processing for internet social images. *Multimedia Tools and Application*, 51:723–738, 2010.

[27] D. Liu, S. Yan, Y. Rui, and H. J. Zhang. Unified tag analysis with multi-edge graph. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[28] X. Liu, B. Cheng, S. Yan, J. Tang, T. C. Chua, and H. Jin. Label to region by bi-layer sparsify priors. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[29] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huang, and H. Jin. Nonparametric label-to-region by search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[30] Y. Liu, T. Mei, and X.-S. Hua. Crowdreranking: Exploring multiple search engines for visual search reranking. In *Proceedings of the ACM SIGIR Conference*, 2009.

[31] T. Mei, Z.-J. Zha, Y. Liu, M. Wang, and et al. Msra at trecvid 2008: High-level feature extraction and automatic search. In *Proceedings of NIST TRECVID workshop*, 2008.

[32] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 2010.

[33] G.-J. Qi, C. C. Aggarwal, and T. Huang. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the International Conference on World Wide Web*, 2011.

[34] M. Rege, M. Dong, and J. Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient

web image clustering. In *Proceedings of the International Conference on World Wide Web*, 2008.

[35] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting images databases from the web. In *Proceedings of the International Conference on Computer Vision*, 2007.

[36] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what i watch: using community activeity to understand content. In *Proceedings of the ACM Workshop on Multimedia Information Retrieval*, 2007.

[37] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu. Transfer learning on heterogenous feature spaces via spectral tranformation. In *Proceedings of the International Conference on Data Mining*, 2010.

[38] B. Sigurbjörnsson and R. V. Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of International Conference on World Wide Web*, 2008.

[39] J. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4:12–20, 1995.

[40] R. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28:49–56, 1995.

[41] A. Sun and S. S. Bhowmick. Quantifying tag representativeness of visual content of social images. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[42] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[43] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceedings of the International Conference on Image and Video Retrieval*, 2008.

[44] G. Wang and D. A. Forsyth. Object image retrieval by exploiting online knowledge resources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[45] J. Wang, Y.-G. Jiang, and S.-F. Chang. Label diagnosis through self tuning for web image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[46] M. Wang, X. S. Hua, R. Hong, J. Tang, G. J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5), 2009.

[47] M. Wang, X. S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3), 2009.

[48] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua. Assistive multimedia tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Survey*, 2011.

[49] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *Proceedings of the ACM Conference on Multimedia*, pages 944–951, 2004.

[50] X.-J. Wang, W.-Y. Ma, L. Zhang, and X. Li. Iteratively clustering web images based on link and attribute reinforcements. In *Proceedings of the ACM Conference on Multimedia*, 2005.

[51] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[52] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized LDA. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[53] R. Yan and A. G. Hauptmann. Co-retrieval: A boosted reranking approach for video retrieval. In *Proceedings of the ACM Conference on Image and Video Retrieval*, 2004.

[54] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proceedings of the ACM Conference on Image and Video Retrieval*, 2003.

[55] K. Yang, X.-S. Hua, M. Wang, and H. C. Zhang. Tagging tags. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[56] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning from image clustering via the social web. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL*, 2009.

[57] Y.-H. Yang, P. Wu, C. W. Lee, K. H. Lin, W. Hsu, and H. H. Chen. Contextseer: Context search and recommendation at query time for shared consumer photos. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[58] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[59] Z.-J. Zha, T. Mei, J. Wang, X.-S. Hua, and Z. Wang. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 2009.

[60] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua. Interactive video indexing with statistical active learning. *IEEE Transactions on Multimedia*, 2011.

[61] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Viusal query suggestion. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[62] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *Proceedings of the International Conference on Computer Vision*, pages 846–851, 2005.

[63] R. Zhao and W. I. Grosky. Narrowing the semantic gap - improved text-based web document retireval using visual fetures. *IEEE Transactions on Multimedia*, 4, 2002.

[64] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[65] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous transfer learning for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.