

# Generating Recommendation Evidence Using Translation Model

Jizhou Huang<sup>†,‡,\*</sup>, Shiqi Zhao<sup>‡</sup>, Shiqiang Ding<sup>‡</sup>, Haiyang Wu<sup>‡</sup>, Mingming Sun<sup>‡</sup>, Haifeng Wang<sup>‡</sup>

<sup>†</sup>Harbin Institute of Technology, Harbin, China

<sup>‡</sup>Baidu Inc., Beijing, China

{huangjizhou01, zhaoshiqi, dingshiqiang01, wuhaiyang, sunmingming01, wanghaifeng}@baidu.com

## Abstract

Entity recommendation, providing entity suggestions relevant to the query that a user is searching for, has become a key feature of today’s web search engine. Despite the fact that related entities are relevant to users’ search queries, sometimes users cannot easily understand the recommended entities without evidences. This paper proposes a statistical model consisting of four sub-models to generate evidences for entities, which can help users better understand each recommended entity, and figure out the connections between the recommended entities and a given query. The experiments show that our method is domain independent, and can generate catchy and interesting evidences in the application of entity recommendation.

## 1 Introduction

Over the past few years, major commercial web search engines have enriched and improved users’ experiences of information retrieval by presenting recommended entities related to their search queries besides the regular search results. Figure 1 shows an example of Baidu (www.baidu.com) web search engine’s recommendation results of the query “Obama”. On the panel, a ranked list of celebrities related to “Obama” is presented, providing users a quick access to entities closely attached to their interests and enhance their information discovery experiences.

Related entity recommendations can increase users’ engagements by evoking their interests to these entities and thereby extending their search sessions [Aggarwal *et al.*, 2015]. However, if users have no background knowledge on a recommended entity, they may possibly be confused, and leave without exploring these entities. In order to help users to quickly understand whether, and why, the entities can meet their interests, it is important to provide evidences on the recommended entities. As depicted in Figure 1, the phrase under each entity, which we refer to as the “recommendation evidence” (“evidence” for short later in this paper), is presented, providing users a quick overview of the representa-

## People also search for



Figure 1: An example of Baidu web search engine’s recommendations for the query “Obama”. The evidences are presented under each entity.

tive features of each entity.<sup>1</sup> Using the third recommended entity “Oh Han Ma” as an example, its evidence is “Obama’s Korean name”. Without this evidence, users may get confused and think it is an unrelated recommendation, because this entity is unfamiliar. Therefore, presenting evidence for recommendation can help in building and promoting trust between the users and the recommendation system [Voskarides *et al.*, 2015]. In this paper, we will focus on generating evidences from sentences. Selecting appropriate evidences for each recommended entity related to a search query will be left for future work.

Although there is some previous work on entity recommendation systems, e.g., [Blanco *et al.*, 2013; Yu *et al.*, 2014; Bi *et al.*, 2015], the problem of providing evidence for recommendation has not been well addressed. To the best of our knowledge, little research has been published on automatically generating evidences for recommended entities. Although extracting evidences from structured data is a feasible way, the coverage is insufficient. Take “Obama” as an example, the disambiguation texts of entities in Wikipedia, e.g., Obama<sup>2</sup>, providing useful and key information to help disambiguate different mentions of an entity, can be used as evidences for recommendation. For example, from the following disambiguation texts of Obama (EXAMPLE 1 and 2), we

<sup>1</sup>We translate both Chinese entities and evidences into English to make it more understandable.

<sup>2</sup>[https://en.wikipedia.org/wiki/Obama\\_\(disambiguation\)](https://en.wikipedia.org/wiki/Obama_(disambiguation))

\*Corresponding author.

could extract “the highest point in Antigua and Barbuda” as an evidence of the entity “Mount Obama” and “the 44<sup>th</sup> and current President of the United States” for “Barack Obama”. But the method does not work for the entity “Oh Han Ma” since no such information is available in Wikipedia.

**EXAMPLE 1.** *Barack Obama (born 1961) is the 44th and current President of the United States*

**EXAMPLE 2.** *Mount Obama, the highest point in Antigua and Barbuda*

We also investigated the largest Chinese online encyclopedia Baidu Baike<sup>3</sup>, to examine the coverage of disambiguation texts in Chinese. Results show that only 3.4% of entities in Wikipedia and 5.9% of entities in Baike have such disambiguation information, which shows that directly extracting evidences from such online resources is not sufficient. In order to provide users with consistent user experiences, it is important to generate evidences for all recommended entities.

In this paper, we propose to use statistical machine translation (SMT) techniques to generate evidences for the entities recommended in web search. We train a translation model on a query-title aligned corpus, derived from clickthrough data of Baidu web search engine. Two additional feature functions are introduced in the translation model to produce attractive evidences.

The major contributions of this paper are summarized as follows:

1. We study the novel issue of evidence generation (EG) for entity recommendation.
2. We propose an SMT based approach to generate evidence candidates for entities, and introduce two additional feature functions in the model to produce attractive evidences. The experimental results show that our approach is very promising.
3. Large-scale monolingual parallel data is essential for training EG models. We propose an efficient method to mine aligned sentence-evidence pairs from the click-through data of a search engine.

## 2 Problem Statement

In this paper, we generate recommendation evidences for an entity from sentences that contain the entity. In order to provide users a quick overview of the representative features of a given entity, we define entity evidence as follows: (1) the evidence must correctly describe the entity; (2) the evidence must be concise so as to be presented in a limited space<sup>4</sup>; (3) the evidence should be informative and attractive so as to attract users to browse and click the recommended entity.

From the definition, we can see that the EG task requires to shorten a sentence by deleting some less important words and/or replacing some phrases with other more concise and attractive phrases, and organize the generated evidences in an accurate, fluent, and catchy manner. An example is depicted in Figure 2.

<sup>3</sup><http://baike.baidu.com/>

<sup>4</sup>In this paper, we constrain that the evidence should be no longer than 10 Chinese characters.

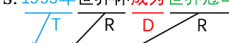
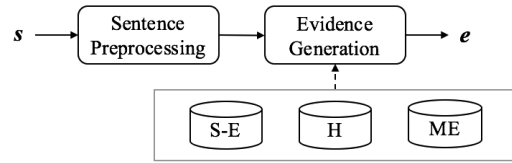
S: 1995年世界杯成为世界冠军 (won the championship at the World Cup in 1995)  
  
 E: 前世界杯世界冠军 (a former World Cup winner)  
 where S = Sentence, E = Evidence, R = Reserve, D = Delete, T = Substitute

Figure 2: An example of evidence generation.



where S-E = S-E Pairs, H = Headlines, ME = Manually-labeled Evidences

Figure 3: Overview of the EG method.

Given a sentence  $s$ , our task is to generate a ranked list of evidences  $E = \{e_1, e_2, \dots, e_k\}$  for  $s$ .

Figure 3 shows an overview of our method. The EG method contains two components, i.e., sentence preprocessing and evidence generation. Sentence preprocessing mainly includes Chinese word segmentation [Peng *et al.*, 2004], POS tagging [Gimenez and Marquez, 2004] and dependency parsing [McDonald *et al.*, 2006] for the input sentences, as POS tags and dependency information are necessary for the following stages. Evidence generation (described in Section 3.2) is designed to generate evidences for the input sentences with a statistical machine translation model. The evidence generation model needs three data sources. Firstly, sentence-evidence parallel corpus (S-E Pairs) is used to train the “translation” model and language model (described in Section 3.1 and 3.2). Furthermore, headlines of news articles (Headlines) and manually-labeled evidences are used to train an attraction model (described in Section 3.2 and 3.3) to increase the attraction of generated evidences.

## 3 Evidence Generation Model

Despite the similarity between evidence generation (EG) and machine translation (MT), the statistical model used in SMT cannot be directly applied in EG, since there are some clear differences between them: (1) the bilingual parallel data for SMT are easy to collect. In contrast, the large monolingual parallel data for EG are difficult to acquire; (2) SMT has a unique purpose, i.e., producing high-quality translations for the inputs. In comparison, EG aims at generating attractive and concise descriptions as evidences; (3) In SMT, there is not any limitation for the length of translations, whereas in EG, the length of evidence is strictly limited due to the layout limitation.

### 3.1 Sentence-evidence Parallel Data

To train the EG model, we need large-scale monolingual parallel data of sentence-evidence aligned pairs. Take the entity “Ilham Anas” in Figure 1 as an example, Table 1 shows several evidences (E) of this entity and their aligned sentences (S) from which the evidences are generated.

<b>S:</b> Ilham Anas is Indonesia’s Obama look alike
<b>E:</b> the Obama’s Indonesian look-alike
<b>S:</b> Ilham Anas shares a striking resemblance with US President Barack Obama
<b>E:</b> a striking resemblance to Obama

Table 1: Examples of aligned sentence-evidence pairs.

	U	U+B	U+B+P	U+B+P+D
Precision	87.9%	88.4%	90.0%	91.6%
Recall	90.2%	90.6%	90.8%	91.2%

Table 2: Performance of evidence classifier.

From the examples shown in Table 1, we can see that the evidences and sentences may use different language styles and vocabularies, so the generation model is required to bridge the gap between them. We view the title-query pairs derived from the clickthrough data of a search engine as sentence-evidence pairs, and use this data to construct the training corpus for evidence generation model, the reasons are as follows:

1. The queries and evidences have similar language styles and vocabularies, because the perplexity result of the language model trained on 12 million random queries tested on 200,000 sample evidences is 578, which indicates high language similarity between them according to the metric described in [Gao *et al.*, 2010].
2. A query is parallel to the titles of documents clicked on for that query [Gao *et al.*, 2010], thus the query-title aligned corpus is a good source of monolingual parallel data.

The task of evidence identification from query logs can be viewed as a binary classification problem of distinguishing evidence from non-evidence. To construct the data for training the classifier, we collected the disambiguation texts of entities from Baidu Baike as seed evidences, and got 488,001 evidences in this way. To enrich the evidence data, we used pattern based method [Muslea, 1999] to extract more evidences from the main texts in Baidu Baike using the pattern “<entity> is <evidence>”. Altogether, we collected 1,040,897 evidences as positive instances. We then extracted equivalently random queries as negative instances with the same length range as the positive instances. To construct the test set for the classifier, we randomly sampled 10% of the data from positive and negative instances separately, and the 90% of data were left for training.

Maximum Entropy is selected as the classification model because it is an effective technique for text classification [Nigam *et al.*, 1999]. The features used for training the evidence classifier are unigrams (U), bigrams (B), POS tagging (P) and dependency parsing (D). Table 2 shows the performance of the evidence classifier.

To extract candidate sentence-evidence pairs, we used the evidence classifier to identify evidences from queries on a six-months clickthrough data of Baidu web search engine. To enrich the data, each title was segmented into multiple sub-titles

using punctuations, and formed multiple title-query pairs. Finally, similar to [Quirk *et al.*, 2004], the pairs were filtered if they met any of the following rules:

- Title and query have no word overlapping;
- Title and query are identical;
- The length<sup>5</sup> of a query is greater than 10 or less than 6;
- Title-query pairs with significantly different lengths (the shorter is less than two-thirds the length of the longer).

A total of 55,149,076 title-query aligned pairs were obtained, which we used as sentence-evidence parallel corpus to train our evidence generation model. Mean edit distance [Levenshtein, 1966] over Chinese characters was 5.7; mean lengths of sentence and evidence were 10.7 and 7.9, respectively.

### 3.2 Evidence Generation Model

Our EG model contains four sub-models: a translation model, a language model, a length model, and an attraction model, which control the adequacy, fluency, length, and attraction of the evidences, respectively.<sup>6</sup>

#### Translation Model (M1)

Evidence generation is a decoding process. Similar to [Zhao *et al.*, 2009], the input sentence  $s$  is first segmented into a sequence of units  $\bar{s}_1^l$ , which are then “translated” to a sequence of units  $\bar{e}_1^l$ . Let  $(\bar{s}_i, \bar{e}_i)$  be a pair of translation units, their translation likelihood is computed using a score function  $\phi_{tm}(\bar{s}_i, \bar{e}_i)$ . Thus the translation score between  $s$  and  $e$  is decomposed into:

$$p_{tm}(\bar{s}_1^l, \bar{e}_1^l) = \prod_{i=1}^l \phi_{tm}(\bar{s}_i, \bar{e}_i)^{\lambda_{tm}}$$

where  $\lambda_{tm}$  is the weight for the translation model. Actually, it is defined similarly to the translation model in SMT [Koehn *et al.*, 2003].

#### Language Model (M2)

We use a tri-gram language model in this work. The language model based score for the evidence  $e$  is computed as:

$$p_{lm}(e) = \prod_{j=1}^J p(e_j | e_{j-2} e_{j-1})^{\lambda_{lm}}$$

where  $J$  is the number of words of  $e$ ,  $e_j$  is the  $j$ -th word of  $e$ , and  $\lambda_{lm}$  is the weight for the language model.

#### Length Model (M3)

We use a length-penalty function to generate short evidences whenever possible. To meet the requirement that the evidence should be less than 10 Chinese characters, the length score for the evidence  $e$  is computed as:

$$p_{lf}(e) = \begin{cases} N & \text{if } N \leq 10 \\ \frac{1}{N-10} & \text{if } N > 10 \end{cases}$$

where  $N$  is the number of Chinese characters of  $e$ .

<sup>5</sup>Throughout this paper the length of sentences, evidences, and queries is assumed to be the number of Chinese characters in them.

<sup>6</sup>The EG model applies monotone decoding, which does not contain a reordering sub-model that is often used in SMT.

### Attraction Model (M4)

The attraction model prefers evidences that can better achieve the requirements of entity recommendation described in Section 2. After analyzing a set of manually labeled entity evidences, we found that the attraction of  $e$  depends on three aspects: the vocabulary used, the language style, and sentence structure. We use two sub-models to capture these aspects. The first one is a special language model trained on headlines of news articles (M4-1 for short), which tries to generate catchy and interesting evidences with similar vocabularies and styles to headlines. The motivation is that news editors usually try their best to use the attractive expressions to write the headlines. The second one is a sentence structure model trained on human annotated evidences (M4-2), which tries to generate evidences with popular syntax styles that users might prefer. Hence the attraction model is decomposed into:

$$p_{am}(e) = p_{hl}(e)^{\lambda_{hl}} \cdot p_{ss}(e)^{\lambda_{ss}}$$

where  $p_{hl}(e)$  is the headline language model and  $p_{ss}(e)$  is the sentence structure model.  $p_{hl}(e)$  is similar to  $p_{lm}(e)$ , but trained on headlines.  $p_{ss}(e)$  is computed as:

$$p_{ss}(e) = \max(K(T_e, T_{t_i}))$$

where  $T_x$  is the dependency tree of sentence  $x$ ,  $t_i$  is the human annotated evidences, and  $K(\cdot, \cdot)$  is the dependency tree kernels described in [Culotta and Sorensen, 2004], which measures the structure similarity between sentences. We combine the four sub-models based on a log-linear framework and get the EG model:

$$\begin{aligned} p(e|s) &= \lambda_{tm} \sum_{i=1}^l \log \phi_{tm}(\bar{s}_i, \bar{e}_i) \\ &+ \lambda_{lm} \sum_{j=1}^J \log p(e_j | e_{j-2} e_{j-1}) + \lambda_{lf} \log p_{lf}(e) \\ &+ \lambda_{hl} \sum_{i=1}^L \log p(e_i | e_{i-2} e_{i-1}) + \lambda_{ss} \log p_{ss}(e) \end{aligned}$$

### 3.3 Resources for Training Attraction Model

To train the attraction model, we need data of headlines and human annotated evidences. We firstly extracted all headlines from three major Chinese news Websites<sup>7</sup>, then we ranked the headlines by the click count in the query logs, and finally top ranked 10 million headlines were remained. To guide the EG model to generate evidences with similar structures to human composed evidences, we need a set of human annotated evidences of high-quality. We used crowdsourcing method [Hsueh *et al.*, 2009] to collect this set. We asked annotators to compose evidences for each sentence, then asked 5 different annotators to vote each evidence with two options: acceptable or not, finally an evidence voted by more than 4 annotators out of 5 as acceptable were kept. A total of 104,775 excellent evidences were obtained.

### 3.4 Parameter Estimation

To estimate parameters  $\lambda_{tm}$ ,  $\lambda_{lm}$ ,  $\lambda_{lf}$ ,  $\lambda_{hl}$ , and  $\lambda_{ss}$ , we adopt the approach of minimum error rate training (MERT) that is popular in SMT [Och, 2003]. In SMT, the optimization

<sup>7</sup>(1) <http://news.qq.com/>, (2) <http://news.sina.com.cn/>, and (3) <http://news.sohu.com/>

ID	Category	Percentage
C1	People	31.6%
C2	Terminology	8.7%
C3	Organization	8.2%
C4	Animal	6.3%
C5	Place	5.9%
C6	Movie	4.8%
C7	Others	34.5%

Table 3: Categories of test entities.

objective function in MERT is usually BLEU [Papineni *et al.*, 2002], which requires human references. To provide human annotated evidences as references for each sentence, we asked 5 annotators to compose evidences for each sentence separately. Finally, we invited other 3 judges to vote each evidence, and evidences with at least two agreements were kept. A total of 7,822 sentences with human annotated evidences were obtained, and only the first evidence of a sentence was used as the reference. We estimate parameters for each model separately. The parameters that result in the highest BLEU score on the development set were finally selected.

## 4 Experimental Setup

We use the method proposed in [Che *et al.*, 2015] as baseline, which used a CRF model to compress sentences by dropping certain less important words. For the EG method proposed in this paper, we have trained three models. The first EG model combines M1, M2, and M3, which is used for evaluating the performance of default features (named as EG-D). The second EG model combines M1, M2, M3, and M4-1, which is used to examine if headlines could help increase the performance (named as EG-H). The third considers all sub-models M1, M2, M3, and M4 (named as EG-F).

### 4.1 Experimental Data

Our method is not restricted in domain or language, since the translation models and features employed here are language independent. Thus sentences in different languages or containing entities of different categories can be used for testing. In this paper, all EG models are trained on a Chinese corpus. Furthermore, to evaluate if our method can generate evidences for sentences of different lengths and categories, in our experiments, we manually select 1,000 Chinese sentences as a test set to carry out the evaluation according to the following rules: (1) the sentence contains descriptive information about an entity in a randomly selected entity set, and (2) the length of a sentence is in the range 5 to 20. The average length of the sentences is 11.8. Finally, to check if our method can generate evidences for different types of entities, we classify the entities described by the collected sentences. After classification, 104 categories in total are obtained. Table 3 shows the percentage of the classification results, in which, the top 6 categories are listed, and all the other categories are combined into ‘‘Others’’.

## 4.2 Evaluation Metrics

The evaluation metrics for EG are similar to the human evaluation for MT [Callison-Burch *et al.*, 2007]. The generated evidences are manually evaluated based on three criteria, i.e., adequacy, fluency, and attraction, each of which has three scales from 1 to 3. Here is a brief description of the different scales for the criteria:

- Adequacy** 1: The meaning is obviously changed.  
2: The meaning is generally preserved.  
3: The meaning is completely preserved.
- Fluency** 1: The evidence  $e$  is incomprehensible.  
2:  $e$  is comprehensible.  
3:  $e$  is a flawless phrase.
- Attraction** 1: The attraction is obviously decreased.  
2: The attraction is generally retained.  
3: The attraction is increased.

To make the attraction understood consistently by raters in practice, we define attraction in detail by using three aspects: the evidence should be more concise, informative, and/or interesting than the sentence.

BLEU is widely used for automatic evaluation in MT. It measures the similarity between a translation and human references. To further assess the quality of the generated evidences, we compute BLEU scores of each method, and 3 human references for each test sentence are provided.

## 5 Results and Analysis

We use the baseline and the three EG models to generate evidences. Results show that the percentages of test sentences that can be generated (“coverage” for short later in this paper) are 99.9%, 88.8%, 87.3%, and 87.3% for baseline, EG-D, EG-H, and EG-F, respectively. The coverage of baseline is much higher, because it is easy to delete words to match the required length without considering other constraints of evidence. The reason why the last two coverages are lower than the second one is that, after adding the sub-models into the EG-D, several extremely long sentences cannot find properly short phrase replacements or do phrase deletion so that the method fails to generate evidences within maximum length allowed. It indicates that generating evidences for extremely long sentences is more difficult than the shorter ones. In our experiments, the first evidence generated by each model is used for evaluation.

### 5.1 Evaluation

We ask two raters to label the evidences generated by baseline and the three models based on the criteria defined in Section 4.2. The labeling results averaged between two raters are shown in Table 5. We can see that for adequacy, fluency, and attraction, the EG-F model gets the highest scores. The percentages of label “3” are 50.9%, 69.9% for adequacy and fluency, which is promising for our EG task. But the percentage of label “3” for attraction is 17.6%, the main reason is that it is difficult to increase much attraction due to the strict length limitation of EG task. This motivates us to further improve the attraction model in the future work.

We compute the kappa statistic between the raters. Kappa is defined as  $K = \frac{P(A) - P(E)}{1 - P(E)}$  [Carletta, 1996], where  $P(A)$

<b>S:</b> 一种特殊的极具价值的体育运动方式 (a special and valuable way of doing sports)
<b>E (Baseline):</b> 一种特殊的极具价值的 (a special and valuable) <sup>‡</sup>
<b>E (EG-D):</b> 一种有价值体育运动 (a valuable sport)
<b>E (EG-H/EG-F):</b> 特殊的有价值体育运动 (physical exercise with special value)
<b>S:</b> 最受企业界人士欢迎十大名嘴之一 (one of the top 10 famous anchormen popular with the business circle)
<b>E (Baseline):</b> 最受企业界人士名嘴 (most by business circle famous anchorman) <sup>‡</sup>
<b>E (EG-D):</b> 最受企业界十大名嘴 (the top 10 famous anchormen by the business world) <sup>‡</sup>
<b>E (EG-H/EG-F):</b> 最受企业人士欢迎名嘴 (the famous anchorman most popular with businessmen)

Table 4: The generated evidences of some sentences. <sup>‡</sup> indicates that the evidence has grammatical errors.

is the proportion of times that the labels agree, and  $P(E)$  is the proportion of times that they may agree by chance. We define  $P(E)=1/3$ , as the labeling is based on three point scales. The results show that the kappa statistics for adequacy, fluency, and attraction are 0.6598, 0.6833, and 0.6763, respectively, which indicates a substantial agreement ( $K$ : 0.61-0.8) according to [Landis and Koch, 1977].

Table 4 shows an example of the generated evidences. Evidences E of baseline, EG-D, EG-H, and EG-F are listed with their source sentences S.

### 5.2 Comparison

We tune the parameters for the three EG models using the development data as described in Section 3.4 and evaluate them with the test data as described in Section 4.1.

As can be seen from the test results in Table 5, the EG-H and EG-F models significantly outperform baseline and EG-D in both human and automated evaluation. Although the coverage of EG-F (87.3%) is lower compared to baseline (99.9%), the usability of EG-F is much higher than that of baseline: (1) the BLEU score is improved by 15.01 (from 53.63 to 68.64), and (2) the overall improvements of labels “2” and “3” are higher for adequacy, fluency, and attraction. Compared with baseline, the EG-F achieves a better balance between coverage and usability. The baseline is not readily applicable to the application of entity recommendation due to its low usability. The overall percentages of labels “2” and “3” of EG-D, baseline, EG-H, and EG-F, in all three evaluation metrics, are largely consistent with movements in BLEU scores, which verifies that BLEU tracks human evaluation well.

As shown in Table 5, the EG-H model outperforms the EG-D model with noticeable improvements, as the percentages of labels “2” and “3” are much higher for all three evaluation metrics. This shows that the model M4-1 can contribute to generating evidences of higher quality. Table 5 also shows that the EG-F model improves the performance compared with EG-H model. This verifies the effectiveness of bringing

		Baseline	EG-D	EG-H	EG-F
<b>Adequacy (%)</b>	<b>1</b>	44.3	45.4	25.5	24.8
	<b>2</b>	22.1	27.0	24.2	24.3
	<b>3</b>	33.6	27.6	50.3	50.9
<b>Fluency (%)</b>	<b>1</b>	25.6	31.9	17.7	17.6
	<b>2</b>	29.4	14.4	13.3	12.5
	<b>3</b>	45.0	53.7	69.0	69.9
<b>Attraction (%)</b>	<b>1</b>	50.2	51.9	30.4	29.3
	<b>2</b>	48.6	38.5	52.5	53.1
	<b>3</b>	1.2	9.6	17.1	17.6
<b>BLEU</b>		53.63	40.59	67.39	68.64

Table 5: The evaluation results of baseline and EG models.

		C1	C2	C3	C4	C5	C6	C7
<b>A</b>	<b>1</b>	21.5	34.8	38.2	22.3	24.5	25.0	22.4
	<b>2</b>	22.2	27.9	28.5	22.3	14.2	20.0	27.2
	<b>3</b>	56.3	37.3	33.3	55.4	61.3	55.0	50.4
<b>F</b>	<b>1</b>	15.3	23.4	23.6	18.7	17.0	12.5	17.4
	<b>2</b>	9.0	18.4	11.8	17.0	11.3	21.2	12.8
	<b>3</b>	75.7	58.2	64.6	64.3	71.7	66.3	69.8
<b>T</b>	<b>1</b>	24.0	39.9	43.7	31.3	25.5	27.5	28.8
	<b>2</b>	59.0	43.7	39.6	46.4	55.6	56.2	53.3
	<b>3</b>	17.0	16.4	16.7	22.3	18.9	16.3	17.9

Table 6: The evaluation results of each category of the EG-F model. In which, A = Adequacy (%), F = Fluency (%), and T = Attraction (%). C1 to C7 are defined in Table 3.

model M4-2 with human annotated data set. It also indicates that more human annotated evidences can be adopted to better guide the EG model to generate evidences similar to human composed evidences.

We further compare the phrase replacements and deletions performed by each model. Experimental results show that the average number of phrase replacements/deletions in sentences are 0/2.2, 0.8/1.9, 0.5/1.7, and 0.5/1.6 for baseline, EG-D, EG-H, and EG-F, and the average lengths of evidences generated by these models are 7.9, 7.2, 7.7, and 7.8, respectively. The baseline conducts more deletions than the other three models, but makes no replacement. As we can see from Table 5, the adequacy, fluency, and attraction of baseline and EG-D are lower than that of EG-H and EG-F models, which demonstrates that some key phrase replacements or deletions are inadequately or incorrectly performed in baseline and EG-D. This motivates us that more efficient models for replacing and deleting can be explored in future to get better results.

Table 6 shows the evaluation results of each category involved in our experiments. Except C2 (terminology) and C3 (organization), the performance of all the other categories match up with or outperform the overall performance of EG-F (shown in Table 5). This verifies that EG-F can generate domain independent evidences which achieve our applications. The percentages of label “1” of C2 and C3 for all three evaluation metrics are higher than that of the overall performance of EG-F. The main reason is that the average lengths of sentences of C2 and C3 are 12.1 and 13.3, which are larger than the overall average length 11.8. Another reason is the “mis-

translation” of the infrequent numbers that occur more frequently in C2 and C3 (account for 24.2% of bad cases), e.g., the “ranked 27th” is wrongly “translated” to “ranked 1st”. It is thus more necessary to bring a new model to “translate” numbers in correct ways to improve the EG performance.

## 6 Related Work

In this section, we review the related work. A research topic closely related to our work is the task of mining evidences for entities. [Li *et al.*, 2013] proposed a method to mine evidences for named entity disambiguation task. The evidences consist of multiple words related to an entity. Our work is different in that we aim at generating comprehensible and human-readable sentences as evidences. Our work is also related to the task of sentence compression. [Turner and Charniak, 2005; Galley and McKeown, 2007; Nomoto, 2007; Che *et al.*, 2015] proposed methods to compress an original sentence by deleting words or constituents. However, these extractive methods are restricted to word deletion, and therefore are not readily applicable to the more complex EG task. [Cohn and Lapata, 2013] proposed an abstractive method to compress an original sentence by reordering, substituting, inserting, and removing its words. This method cannot be directly transplanted to the EG task due to the specificity of the entity evidence, since the EG task requires to generate evidences within specified length limits by using attractive expressions and vocabularies from a sentence, rather than simply compress a sentence.

Our work is also closely related to the studies in sentential paraphrase generation using monolingual machine translation. Although the studies share the same idea in translating a source sentence into a target sentence that are in the same language by using monolingual parallel corpus, there are some differences from our work. [Wubben *et al.*, 2012] built a monolingual machine translation system to convert complex sentences into their simpler variants. While our work aims at generating concise, informative, and interesting evidences from sentences rather than just simplifying sentences. [Quirk *et al.*, 2004; Zhao *et al.*, 2009] viewed paraphrase generation as monolingual machine translation, which aims to generate a paraphrase for a source sentence in a certain application. The three major distinctions between evidence generation and these studies are: (1) we consider language styles and vocabularies in evidences; (2) we add a length model to ensure the generating of evidences within maximum length allowed; (3) we introduce two attraction measures and features in the EG model, to produce more attractive evidences from sentences.

## 7 Conclusion and Future Work

In this paper, we study the problem of generating evidences for the recommended entities in web search. We propose a translation model to generate evidences from sentences. The experiments show that our method can generate domain independent evidences with high usability.

As future work, we plan to dynamically select appropriate evidences for each recommended entity related to the search query. We are interested in generating evidence using multiple sentences, rather than relying on a single sentence.

## Acknowledgments

This research is supported by the National Basic Research Program of China (973 program No. 2014CB340505). We would like to thank the anonymous reviewers for their insightful comments.

## References

- [Aggarwal *et al.*, 2015] Nitish Aggarwal, Peter Mika, Roi Blanco, and Paul Buitelaar. Insights into entity recommendation in web search. In *ISWC*, 2015.
- [Bi *et al.*, 2015] Bin Bi, Hao Ma, Bo-june Paul Hsu, Wei Chu, Kuansan Wang, and Junghoo Cho. Learning to recommend related entities to search users. In *WSDM*, pages 139–148. ACM Press, 2015.
- [Blanco *et al.*, 2013] Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, and Nicolas Torzec. Entity recommendations in web search. In *ISWC*, pages 33–48, 2013.
- [Callison-Burch *et al.*, 2007] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of machine translation. In *ACL*, pages 136–158, 2007.
- [Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, pages 249–254, 1996.
- [Che *et al.*, 2015] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. Sentence compression for aspect-based sentiment analysis. *Audio, Speech, and Language Processing*, 23(12):2111–2124, 2015.
- [Cohn and Lapata, 2013] Trevor Cohn and Mirella Lapata. An abstractive approach to sentence compression. *ACM Trans. Intell. Syst. Technol. (TIST)*, 4(3), 2013.
- [Culotta and Sorensen, 2004] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *ACL*, pages 423–429, 2004.
- [Galley and McKeown, 2007] Michel Galley and Kathleen McKeown. Lexicalized Markov grammars for sentence compression. In *HLT-NAACL*, pages 180–187, 2007.
- [Gao *et al.*, 2010] Jianfeng Gao, Xiaodong He, and Jianyun Nie. Clickthrough-based translation models for web search: from word models to phrase models. In *CIKM*, pages 1139–1148. ACM Press, 2010.
- [Gimenez and Marquez, 2004] Jesus Gimenez and Lluís Marquez. SVMTool: A general POS tagger generator based on Support Vector Machines. In *LREC*, 2004.
- [Hsueh *et al.*, 2009] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. In *HLT-NAACL*, pages 27–35, 2009.
- [Koehn *et al.*, 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL*, pages 48–54, 2003.
- [Landis and Koch, 1977] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977.
- [Levenshtein, 1966] V Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physice-Doklady*, 10:707–710, 1966.
- [Li *et al.*, 2013] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. Mining evidences for named entity disambiguation. In *KDD*, pages 1070–1078, 2013.
- [McDonald *et al.*, 2006] Ryan McDonald, Kevin Lerman, and Fernando Pereira. Multilingual dependency analysis with a two-stage discriminative parser. In *CoNLL*, pages 216–220, 2006.
- [Muslea, 1999] Ion Muslea. Extraction patterns for information extraction tasks: A survey. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.
- [Nigam *et al.*, 1999] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, 1999.
- [Nomoto, 2007] Tadashi Nomoto. Discriminative sentence compression with conditional random fields. *Information processing & management*, 43(6):1571–1587, 2007.
- [Och, 2003] Franz Josef Och. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167, 2003.
- [Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.
- [Peng *et al.*, 2004] Fuchun Peng, Fangfang Feng, and Andrew Mccallum. Chinese segmentation and new word detection using conditional random fields. In *COLING*, pages 562–568, 2004.
- [Quirk *et al.*, 2004] Chris Quirk, Chris Brockett, and William B Dolan. Monolingual machine translation for paraphrase generation. In *EMNLP*, pages 142–149, 2004.
- [Turner and Charniak, 2005] Jenine Turner and Eugene Charniak. Supervised and unsupervised learning for sentence compression. In *ACL*, pages 290–297, 2005.
- [Voskarides *et al.*, 2015] Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. Learning to explain entity relationships in knowledge graphs. In *ACL-IJCNLP*, pages 564–574, 2015.
- [Wubben *et al.*, 2012] Sander Wubben, Antal van den Bosch, and Emiel Kraemer. Sentence simplification by monolingual machine translation. In *ACL*, pages 1015–1024, 2012.
- [Yu *et al.*, 2014] Xiao Yu, Hao Ma, Bo-june Paul Hsu, and Jiawei Han. On building entity recommender systems using user click log and freebase knowledge. In *WSDM*, pages 263–272, 2014.
- [Zhao *et al.*, 2009] Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. Application-driven statistical paraphrase generation. In *ACL-IJCNLP*, pages 834–842, 2009.