Morph Resolution Based on Autoencoders Combined with Effective Context Information

Jirong You[1,2], Ying Sha[1,2], Qi Liang[1,2], and Bin Wang[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{youjirong, shaying, liangqi, wangbin}@iie.ac.cn,
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. In social networks, people often create morphs, a special type of fake alternative names for avoiding internet censorship or some other purposes. How to resolve these morphs to the entities that they really refer to is very important for natural language processing tasks. Although some methods have been proposed, they do not use the context information of morphs or target entities effectively; only use the information of neighbor words of morphs or target entities. In this paper, we proposed a new approach to resolving morphs based on autoencoders combined with effective context information. First, in order to represent the semantic meanings of morphs or target candidates more precisely, we proposed a method to extract effective context information. Next, by integrating morphs or target candidates and their effective context information into autoencoders, we got the embedding representation of morphs and target candidates. Finally, we ranked target candidates based on similarity measurement of semantic meanings of morphs and target candidates. Thus, our method needs little annotated data, and experimental results demonstrated that our approach can significantly outperform state-of-the-art methods.

Keywords: Morph · Morph Resolution · Effective Context Infomation · Autoencoder.

1 Introduction

In social networks, people often create morphs, a special type of fake alternative names for avoiding internet censorship or some other purposes [9]. Creating morphs is very popular in Chinese social networks, such as Chinese Sina Weibo. As shown in 1, there is a piece of Chinese Sina Weibo tweet. Here a morph "Little Leo" (小李子) was created to refer to "Leonardo Wilhelm DiCaprio" (莱昂纳多) [1]. The term of "Leonardo Wilhelm DiCaprio" is called this morph's target entity.

Morph resolution is very important in Nature Language Processing (NLP) tasks. In NLP, the first thing is to get the true meanings of words, especially

[1] Leonardo Wilhelm DiCaprio is a famous actor.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

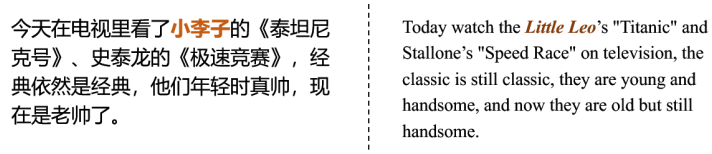

Fig. 1. An example of morph use in Sina Weibo.

including these morphs. Thus, the successful resolution of morphs is the foundation of many NLP tasks, such as word segmentation, text classification, text clustering, and machine translation.

Many approaches are proposed to solve morph resolution. Huang et al. [3] can be considered to have had the first study on this problem, but their method need a large amount of human-annotated data. Zhang et al. [16] proposed an end-to-end context-aware morph resolution system. Sha et al. [9] proposed a framework based on character-word embeddings and radical-character-word embeddings to explore the semantic links between morphs and target entities. These methods do not use the context information of morphs and target entities effectively, only use the context information of neighbor words of morphs and target entities. But there are some neighbor words are unrelated with the semantic links between morphs and target entities. There is still some room for improvement in accuracy of morphs resolution.

In this paper, we proposed a framework based on autoencoders combined with effective context information of morphs and target entities. First, we analyzed what context information are useful for morph resolution, and designed a context information filter to get effective context information by using pointwise mutual information. Second, we proposed a variant of autoencoders which can combine semantic vectors of morphs or target candidates and their effective context information, and we used the combined vectors as the semantic representations of morphs and target candidates respectively. Finally, we ranked target candidates based on similarity measurement of semantic meanings of morphs and target candidates. Using this method, we only take consider of the effective context information of morphs and target entities, and use autoencoders to get essential semantic characteristics of morphs and target entities. Experimental results show that our approach outperforms the state-of-the-art method.

Our paper offers the following contributions:

1. We proposed a new framework based on autoencoders combined with effective context information of morphs and target entities. Our approach outperforms the state-of-the-art method.
2. To get the effective context information of morphs and target entities, we leveraged pointwise mutual information between terms. This helps generate more accurate semantic representation of terms, and can improve the accuracy of morph resolution.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

3. We proposed a variant of autoencoders to generate semantic representation of terms. The autoencoders can combine morphs or target entities and their effective context information and extract essential semantic characteristics of morphs and target entities.

2 Related Work

The study of morphs first appeared in some normalization work on non-formal texts using internet language. For example, Wong et al. [14] examine the phenomenon of word substitution based on phonetic patterns in Chinese chat language, such as replacing "我"(Me, pronounced 'wo') with "偶"(pronounced 'ou'), which is similar to morphs. Early normalization work on non-formal text mainly uses rules-based approaches [14][15][10]. Later, some approaches combine statistics learning with rules to work on the normalization task [13][12][2][5]. Wang et al. [13] establish a probabilistic model based on typical features of non-formal texts including phonetic, abbreviation, replacement, etc., and train it through supervised learning on large corpus.

The concept of morph first appeared in the study of Huang et al. [3]. Huang et al. [3] study the basic features of morphs, including surface features, semantic features and social features. Based on these features, a simple classification model was designed for morph resolution. Zhang et al. [16]also propose an end-to-end system including morph identification and morph resolution. Sha et al.[9] propose a framework based on character-word embedding and radical-character-word embedding to resolve morph after analyzing the common characteristic of morphs and target entities from cross-source corpora. Zhang et al. summarize eight types of patterns of generating morphs, and also study how to generate new morphs automatically based on these patterns [16].

Autoencoders are neural networks capable of learning efficient representations of the input data, without any supervision [11]. Autoencoders can act as powerful feature detectors. There have been many variations of autoencoders. The context-sensitive autoencoders[1] integrate context information into autoencoders and obtain the joint encoding of input data. In this paper, we adopted a similar model of context-sensitive autoencoders to get the semantic representation of morphs and target candidates. We don't need to prepare much annotation data since autoencoder is an unsupervised algorithm.

In this paper, aiming at making full use of effective context information of morphs and target entities, we proposed a new framework based on autoencoders combined with extracted effective context information. Compared with the current methods, our approach only incorporates the effective context information of related words, and outperforms the state-of-the-art methods.

3 Problem Formulation

Morph resolution: Given a set of morphs, our goal is to figure out a list of target candidates which are ranked on the probability of being the real target entity.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

Given documents set $D = \{d_1, d_2, \ldots, d_{|D|}\}$, and morphs set $M = \{m_1, m_2, \ldots, m_{|M|}\}$. Each morph m_i in set M and their real target entities are all appeared in documents set D. Our task is to discover a list of target candidates from D for each m_i, and rank the target candidates based on the probability of being the real target entity.

As shown in Figure 1, the morph "Little Leo" (小李子) was created to refer to "Leonardo Wilhelm DiCaprio" (莱昂纳多). Given the morph "Little Leo" and tweets set from the Sina Weibo, our goal is to discover a list target candidates from tweets and rank the target candidates based on the probability of being the real target entity. The word "Leonardo Wilhelm DiCaprio" is expected to the first result (the real target entity) in the ranked target candidates list.

4 Resolving Morphs Based on Autoencoders Combined with Effective Context Information

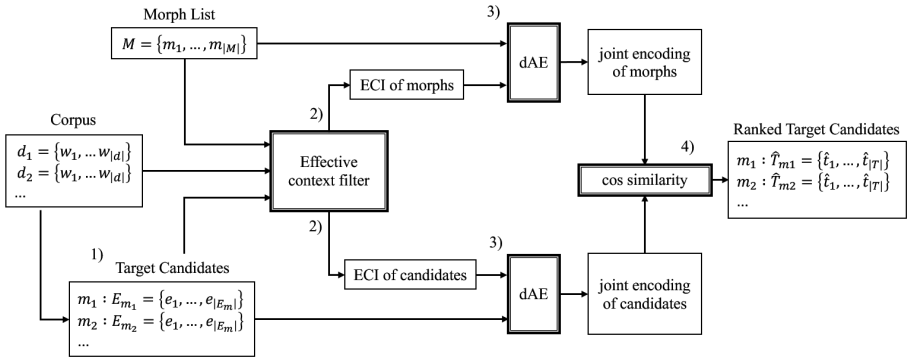

Fig. 2. The Procedure of Morph Resolution.

We designed a framework based on autoencoders combined with effective context information to solve this problem. The procedure of our algorithm is shown in Figure 2. The procedure of morph resolution mainly consists of the following steps:

1. Preprocessing
   In this step, we aim to filter out unrelated terms and extract target candidates $E_{m_i} = \{e_1, e_2, \ldots, e_{|E_{mi}|}\}$. We use two steps to extract the target candidates: 1) tweets which contain morphs are retrieved. Then we can get the published time of these tweets as these morphs'appearing time. Sha et.al discovered that morphs and target entities are highly consistent in temporal distribution [9]. Thus we set a time slot of 4 days to collect tweets which may contain target candidates of morphs; 2) since most morphs refer to named

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

entities, such as the names of persons, organizations, locations, etc. We only need to focus on named entities in these tweets in order to find target candidates of morphs. We can use many off-the-shelf tools working on POS (Part of Speech) and NER (Named Entities Recognition) tasks, including NLPIR[17], Standford NER [6] and so on.

2. Extracting effective context information
   We leverage effective context information (ECI) to generate semantic representation of morphs and target candidates. The effective context information are contextual terms whose semantic relationship with their target term is closer than others. The effective context information can effectively distinguish the characters of the morphs and their targets entities from other terms.
3. Autoencoders combined with effective context information
   We use deep autoencoders (dAE) to get joint encoding representation of morphs or target candidates and their effective context information. Autoencoder can fusion different types of features and embed them into an encoding, which is much more flexible than traditional word embedding methods
4. Ranking target candidates
   After creating encoding representation of morphs and target candidates, we can rank target candidate e_j by calculating cosine similarity between encodings of morph and target candidate. The larger value of cosine similarity between the morph and the target candidate, the more likely the candidate is the real target entity of the morph. The ranked target candidates sequence $\hat{T}_{mi}$ is the result of morph resolution.

In the following sections, we will focus on these two steps: "extracting effective context information" and "autoencoders combined with effective context information".

4.1 Extracting Effective Context Information

To extract the effective context information, we use the pointwise mutual information (PMI) to select right terms that are related with morphs or target entities. PMI is easy to calculate:

$$PMI(x;y) = log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x)$ and $p(y)$ refers to the probability of occurrence of terms x and y in the corpus respectively. $\frac{p(x,y)}{p(x)p(y)}$ represents the co-occurrence of two terms.

PMI quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. PMI maximizes when x and y are perfectly associated (i.e. $p(x,y) = p(x)p(y)$). We use PMI to find collocations and associations between words. Good collocation pairs have high PMI because the probability of co-occurrence

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

is only slightly lower than the probabilities of occurrence of each word. Conversely, a pair of words whose probabilities of occurrence are considerably higher than their probability of co-occurrence gets a small PMI score.

Given a word w, we collect all contextual terms of w from preprocessed tweets which contain w as the set C_w. Note that we also need to remove auxiliary and preposition, since they are useless for our following method. Next, for each term $c_i \in C$, we will calculate PMI between w and c_i, and get the terms of top-k PMI as effective context information of w. In the same way, we can get the effective contextual terms set of all morphs and target candidates.

Terms		Words of top-5 PMI (the value of PMI)				
Morph	The flash (闪电侠) [2]	little emperor (小皇帝)	King James (詹皇)	Bosh (波什)	Dragon King (龙王)	James (詹姆斯)
	PMI	4.406	3.951	3.120	2.884	2.460
Target entity	Wade (韦德)[3]	Bosh (波什)	James (詹姆斯)	King James (詹皇)	little emperor (小皇帝)	Kobe (科比)
	PMI	9.799	9.273	5.856	2.943	2.916
Non-target	Yao (姚明)[4]	Yi Jianlian (易建联)	O'Neill (奥尼尔)	big shark (大鲨鱼)	Lin Shuhao (林书豪)	Howard (霍华德)
	PMI	4.680	4.618	3.488	2.777	2.732
Non-target	Beckham (贝克汉姆)[5]	Giggs (吉格斯)	becky (小贝)	Federer (费德勒)	Olympic Team (国奥队)	Richards (理查兹)
	PMI	4.219	3.398	2.614	2.574	2.526

Table 1. Contextual terms of top-5 PMI of morphs, target entities, and non-target entities.

Table 1 shows contextual terms of top-5 PMI of morphs, target entities, and non-target entities. Here we regard these contextual terms as effective context information. Each row shows the effective contextual terms of different words. The first and second rows show the effective contextual terms of morph "The Flash (闪电侠)"and its target entity "Wade(韦德)"; and the third and fourth rows show the effective contextual terms of non-target entities "Yao(姚明)"and "Beckham(贝克汉姆)". We can discover that the effective contextual terms of the morph and its target entity are nearly consistent, but the effective contextual terms of the morph are completely different from those of non-target entities.

This means that the effective context terms can distinguish the target entity from non-target entities. Effective contextual terms have high PMI with the

[2] A morph of the word 'Wade'.
[3] Dwyane Tyrone Wade, an American professional basketball player for the Cleveland Cavaliers of NBA.
[4] Yao Ming, a Chinese professional basketball player.
[5] David Robert Joseph Beckham, an English former professional footballer.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

morph "The flash" and its target entity "Wade", but have low PMI with those non-target entities like "Yao"(姚明) and "Beckham"(贝克汉姆).

The results mean that we can extract effective contextual terms set by using PMI. Morphs and target entities should have similar context information, so we could extract similar contextual terms set of morphs and target entities by using PMI. Through these contextual terms, we can get more accurate semantic links between morphs and target entities.

4.2 Incorporating Effective Context Information into Autoencoders

In this section, we want to get the representation of essential characters of morphs or target candidates by incorporating effective contextual terms into autoencoders.

Autoencoders neural network is an unsupervised learning algorithm, which encodes its input x into the hidden representation h, then reconstruct x with h precisely:

$$h = g(Wx + b) \tag{2}$$

$$\hat{x} = g(W'h + b') \tag{3}$$

$\hat{x}$ is the reconstruction of x. $W \in R^{d \times d'}, b \in R^{d'}, W' \in R^{d' \times d}$ are the parameters the model learned during training, d and d' means dimension of vectors before and after encoder respectively. Usually $d' \leq d$ for dimensionality reduction. Function g is the activation function in neural network. Figure 3(a) shows the structure of a basic single-layer autoencoder, it can be a cell in deep autoencoders (dAE).

In order to incorporate effective context information into autoencoders, we need to extend the inputs of autoencoders. As Figure 3(b) shown, besides the term w, we also input the effective context information of w. First, we extract C_w^f, the effective context information of w by using methods in 4.1. Second, we generate the word embeddings of each term in C_w^f, and set cc_x as the average of these word embeddings. There are many word embedding methods, such as word2vec[7] or GloVe[8]. Third, we generate u_{cc} , the hidden encoding representation of effective context information by using autoencoders whose input is cc_x. Finally, we can incorporate u_{cc} into deep autoencoders to generate the joint encoding for input terms and their effective context information.

For each layer k^{th} in deep autoencoders, the encoder turns $h_{k-1}(h_0 = x \; if \; k = 1)$ and u_{cc} into one hidden presentation as follows:

$$h = g(W_k h_{k-1} + U_k u_{cc} + b_k) \tag{4}$$

$$\hat{h}_{k-1} = g(W_k' h + b_k') \tag{5}$$

$$\hat{u}_{cc} = g(U_k' h + b_k'') \tag{6}$$

where $\hat{h}_{k-1}$ and $\hat{u}_{cc}$ are the reconstruction of h_{k-1} and u_{cc}. Equation 4 encodes h_{k-1} and u_{cc} into intermediate representation h; and equation 5 and 6 decode h

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

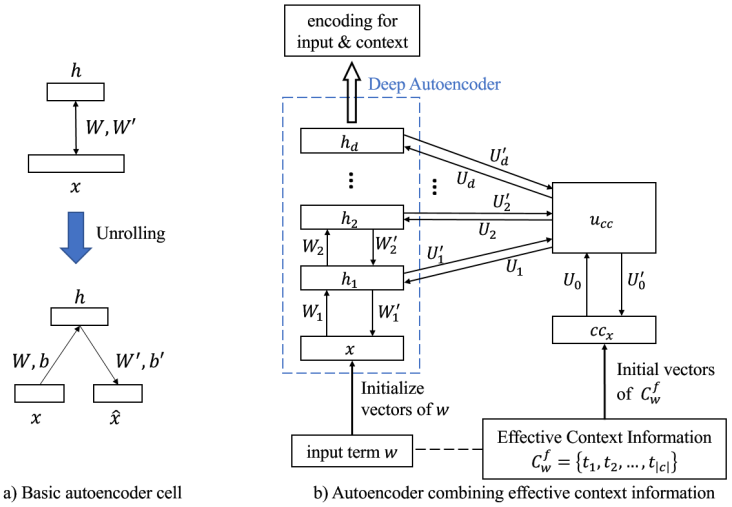

Fig. 3. a) A basic autoencoders cell; b) autoencoders combined with effective context information using PMI filter.

into h_{k-1} and u_{cc}. $W_k, U_k, b_k, W'_k, b'_k, U'_k$, and b''_k are the parameters the model learned during training. The whole model is composed of a stacked set of these layers. The last hidden layer h_d is the joint encoding for input terms and their effective context information.

For the whole model, the loss function must include both deviation of $(x, \hat{x})$ and $(u_{cc}, \hat{u}_{cc})$:

$$loss(x, u_{cc}) = \|x - \hat{x}\|^2 + \lambda \|u_{cc} - \hat{u}_{cc}\|^2 \tag{7}$$

where $\lambda \in [0, 1]$ is the weight that controls the effect of context information during encoding. And the optimize target is to minimize the overall loss:

$$\begin{aligned} &\min_{\Theta} \sum_{i=1}^{n} loss(x^i, u^i_{cc}), \\ &\Theta = \{W_k, W'_k, U_k, U'_k, b_k, b'_k, b''_k\}, \quad k \in 1, 2, ..., depth \end{aligned} \tag{8}$$

we can use back-propagation and the stochastic gradient descent algorithm to learn parameters during training. The autoencoders combined with effective context information is an unsupervised neural network, so we can train the model with a little annotation data.

After training, we can use the autoencoders to generate encoding representations of morphs and target candidates. First, we obtain initial embedding vectors of terms and effective context information, then input these vectors into the autoencoders to obtain the last hidden layer representation, the joint encoding representations of morphs and target candidates respectively. Next, we can rank

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

target candidates by calculating cosine similarity between the joint encodings of morphs and target candidates.

5 Experiments and Analysis

5.1 Datasets

We updated the datasets of Huang's work [3], and added some new morphs and tweets. At last, the datasets include 1,597,416 tweets from Chinese Sina Weibo and 25,003 tweets from Twitter. The time period of these tweets is from May, 2012 to June, 2012 and Sept, 2017 to Oct, 2017. There are 593 pairs of morphs in datasets.

5.2 Parameters Setting

In order to get the appropriate parameters in the model, we randomly selected 50,000 tweets as the verification set to adjust the parameters, including the context window wd, the number of terms for effective context information K of the PMI context filter, and the depth, the encoding dimension $d2$ and λ of the autoencoders. We choose the best parameters after the test on the verification set. During preprocessing, according to Sha's work[9], we set the time window of China Sina Weibo as one day, set the time window of Twitter as three days when we obtain the target candidates. In the initialization of vectors of terms, we choose word2vec[7] to generate 100-dimensional word vectors, which is a popular choice among studies of semantic representation of word embedding. For PMI context filter, we set the context window $wd = 20$, the number of terms for effective context information $K = 100$. For the autoencoder, set the depth as 3, the encoding dimension $d2 = 100$ and $\lambda = 0.5$. In the experiment, we use Adaptive Moment Estimation (Adam) [4] to speedup the convergence rate of SGD.Later we will discuss the effects of different parameters in the task of morph resolution.

5.3 Results

We choose indicator $Presice@K$ to evaluate the result of morph resolution since the result of this task is a ranked sequence. In this paper, $Precise@K = N_k/Q$, means for each morph m_i, if the position of e_{m_i} that is the real target of m_i in result sequence T_m is p, then N_k means the number of resolution results that $p \leq k$, and Q is the total number of morphs for the test. The performance of our approach and some other approaches are presented in Table 2 and Figure 4. *Huang et al.* refers to work in [3], *Zhang et al.* refers to work in [16], *CW* refers to work by Sha et al.[9], while our approach is marked as *AE-ECI*. From the result we can find that our approach outperforms state-of-the-art methods.

The results show that the introduction of effective context information improves the accuracy of morph resolution. The current best method, Sha's work,

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

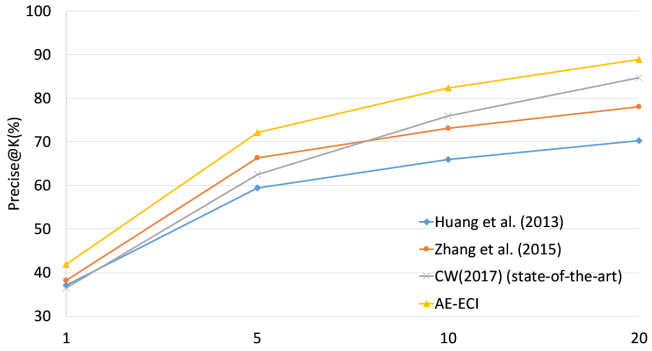

Fig. 4. Performance of several approaches on *pre@k* for morph resolution.

Precise-K (%)	Pre@1	Pre@5	Pre@10	Pre@20
Huang et al.(2013)	37.09	59.40	65.95	70.22
Zhang et al.(2015)	38.17	66.38	73.07	78.06
CW(2017) (state-of-the-art)	36.50	62.50	75.90	84.70
AE-ECI	**41.88**	**72.07**	**82.33**	**88.89**

Table 2. Performance of several approaches on *pre@k* for morph resolution.

just directly uses word embedding to calculate cosine similarity among words. This method only considers context information of neighbor words of morphs or target entities. But there are some neighbor words not having semantic links between morphs or target entities. In our approach, we selects terms that can effectively distinguish the characteristics of target entities from non-target entities by using PMI. Thus we can resolve the morphs more precisely.

5.4 Analysis

In this section, we discuss the effects of different parameters.

Window size and number of context terms. In PMI context filter, we select different window sizes wd and different numbers of contextual terms K to find out impact of window size and number of context terms. However, it seems that wd and K have little impact on performance. The details are shown as Table 3.

wd	5	10	10	20	20	50
K	10	10	20	10	20	10
Pre@1	40.31	41.45	**41.88**	**41.88**	41.73	41.59

Table 3. Effects of Window size and number of context terms.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

Depth and Dimension of Autoencoder. Depth and dimension of autoencoders also have impact on performance. We select different combinations of depth and dimension for experimental verification, and the results show that too large or too small dimension has negative impact on performance. The possible reason may be that the ability of representation of autoencoders with too small dimension is insufficient, while autoencoders with too large dimension is hard to train. The impact of depth is similar. It seems that effect of depth is not very obvious when depth is small; but too deep model performs worse. The details are shown as Table 4.

dimension	50	100	200	300	500	100	100	100	100
depth	3	3	3	3	3	1	2	5	10
Pre@1	39.60	**41.88**	41.59	41.02	39.31	41.59	41.31	41.45	36.89

Table 4. Effects of depth and dimension of autoencoders.

Lambda. λ is the weight that controls effects of effective context information in encoding. We test *Pre@1* of morph resolution at different values of λ. When $\lambda = 0$ it means the effective context information is not added into the model. As shown in Table 5, we find that adding effective context information can improve the performance of model. If λ is too large, it will have negative impact on performance.

λ	0.0	0.1	0.5	1.0
Pre@1	39.88	41.31	**41.88**	40.31

Table 5. Effects of λ.

6 Conclusion

In this paper, we proposed a new approach to solve the problem of morph resolution. By analyzing the features of contextual terms of morphs and their targets, we try to extract effective context information based on PMI. We also proposed autoencoders combined with effective context information to get semantic representations of morphs and target entities. Experimental results demonstrate that our approach outperforms the state-of-the-art work on morph resolution. Next, we will try to extract topic information and integrate it to our models to improve the accuracy of morph resolution, and explore the better ways to fuse the semantic vectors of morphs or target entities and contextual terms.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

7 Acknowledgments

This work is supported by National Science and Technology Major Project under Grant No. 2017YFB0803003, No. 2016QY03D0505 and. No. 2017YFB0803301.

References

1. Amiri, H., Resnik, P., Boyd-Graber, J., Daumé III, H.: Learning text pair similarity with context-sensitive autoencoders. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1882–1892 (2016)
2. Han, B., Cook, P., Baldwin, T.: Automatically constructing a normalisation dictionary for microblogs. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 421–432. Association for Computational Linguistics (2012)
3. Huang, H., Wen, Z., Yu, D., Ji, H., Sun, Y., Han, J., Li, H.: Resolving entity morphs in censored data. In: ACL (1). pp. 1083–1093 (2013)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Li, Z., Yarowsky, D.: Mining and modeling relations between formal and informal chinese phrases from web corpora. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1031–1040. Association for Computational Linguistics (2008)
6. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations). pp. 55–60 (2014)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
8. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
9. Sha, Y., Shi, Z., Li, R., Liang, Q., Wang, B.: Resolving entity morphs based on character-word embedding. Procedia Computer Science **108**, 48–57 (2017)
10. Sood, S.O., Antin, J., Churchill, E.F.: Using crowdsourcing to improve profanity detection. In: AAAI Spring Symposium: Wisdom of the Crowd. vol. 12, p. 06 (2012)
11. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research **11**(Dec), 3371–3408 (2010)
12. Wang, A., Kan, M.Y.: Mining informal language from chinese microtext: Joint word recognition and segmentation. In: ACL (1). pp. 731–741 (2013)
13. Wang, A., Kan, M.Y., Andrade, D., Onishi, T., Ishikawa, K.: Chinese informal word normalization: an experimental study. In: IJCNLP (2013)
14. Wong, K.F., Xia, Y.: Normalization of chinese chat language. Language Resources and Evaluation **42**(2), 219–242 (2008)
15. Xia, Y., Wong, K.F., Li, W.: A phonetic-based approach to chinese chat text normalization. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. pp. 993–1000. Association for Computational Linguistics (2006)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44

16. Zhang, B., Huang, H., Pan, X., Li, S., Lin, C.Y., Ji, H., Knight, K., Wen, Z., Sun, Y., Han, J., et al.: Context-aware entity morph decoding. In: ACL (1). pp. 586–595 (2015)
17. Zhou, L., Zhang, D.: Nlpir: A theoretical framework for applying natural language processing to information retrieval. Journal of the Association for Information Science and Technology **54**(2), 115–123 (2003)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_44