Vol. , No. (2018) 1–10

Domain-independent Intent Extraction from Online Texts

Tran Nhu Thuat[1], Nguyen Huu Hong[1], Nguyen Thanh Tung[1], Dang Tien Son[1], Luong Thai Le[1,2], Phan Xuan Hieu[1]

[1]*Faculty of Information Technology, VNU University of Engineering and Technology, 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam*
[2]*Faculty of Information Technology, University of Transport and Communications, 3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Ha Noi, Vietnam*

Abstract

Identifying user's intents from texts on online channels has wide range of applications from entrepreneurship, banking to e-commerce. However, intent identification is not a simple task due to intent and its attributes are various and strongly depend on the domain of data. In our research, we study the problem of domain-independent intent identification from posts and comments crawled from social networks and discussion forums. We present ten general labels, i.e. labels do not depend on a specific domain, and utilize them when extracting intent and its related information. We also propose the map between general labels and domain-specific labels. We extensively conduct experiments to explore the efficiency of using general labels compared to specific labels in extracting user's intents when the number of intent domains increases. Our study is conducted on a medium-sized dataset from three selected domains: *Tourism*, *Real Estate* and *Transportation*. In term of accuracy, when the number of domains grows, our proposal achieves significantly better results than the domain-specific method in identifying user's intent.

Keywords: Information extraction, intent identification, intent mining.

1. Introduction

Internet users nowadays tend to spend more and more time on social media platforms to post and share their needs, their desires and their intentions. Naturally, such data offers great opportunities for the enterprises, services, retailers...to find and meet their potential customers. For instances, this is the post in a forum of the website webtretho.com: *"Nhà tớ đi Đà Nẵng ngày 14/6 đến 18/6, nhà có 5 người lớn và 1 trẻ em (1 tuổi), các bác thông thái tư vấn cho tớ chọn khách sạn và đi tham quan những đâu là hợp lý nhất, thanks"* (Our family are going to go to Da Nang from 14/6 to 18/6, we have 5 adults and 1 child (1 year old), could you recommend us the hotel, the best places to visit there, thanks). And this is another one from website batdongsan.com: *"Tôi muốn mua đất gần khu công nghiệp Yên Phổ. Diện tích khoảng 90-120m2, giá giao động khoảng 1,4 tỷ đổ lại, có thể kinh doanh hoặc làm nhà trọ cho công nhân thuê"* (I want to buy land near Yen Pho industrial zone. The acreage is about 90-120m2, the price is under

1,4 billion, and the land may be used for lease or business). If a travel agent could take the user intent information timely from the first post, they would give the in line advertising strategy to that user. Clearly, this advertising is very effective because it is provided to whom that need it. And the same thing would happen to a real estate agent if they could get the information from the second post.

In our previous paper [7], we proposed the fully intent understanding include three major stages: *"user intent filtering, intent domain identification, and intent parsing and extraction"*. The first phrase helps to filter text posts on online social media channels to determine which posts contain user intents. The second one inturn, will analyze and identify the domain of the intent, such as *"real–estate, finance–banking, tourism–vacation"*. After that, the text post containing an intent and its domain will be sent to the last stage. This stage will parse, analyze, and extract all the information about the intent. But we recognize that if we do the information extraction in each intent domain separately, it will take a lot of time and effort. Specifically, for each domain, we have to collect the data, build the suitable set of labels, tag the data along to those labels, and then train the individual model. So, in this study, we propose a new method to deeply extract the user intentions without the need of the second stage, *"intent domain identification"*. We call it the domain-independent approach for user intent identification. To address this problem, we choose three intent domains to crawl the data and then analyze them, they are: *tourism, real estate, transportation*. The first thing we do is building the set of specific labels for identifying crucial information related to user intents in each domain. Then these sets are aggregated to form the most suitable list of general labels that include 10 tags. We will discuss more clearly about this process below. For building our model, we carefully do the experimental with three state-of-the-art machine learning models for sequence labeling problem, i.e. Conditional Random Fields (CRFs), Bidirectional Long Short-term Memory (Bi-LSTM) and Bidirectional Long Short-term Memory combined with Conditional Random Fields (Bi-LSTM-CRFs) to make the comparison. Furthermore, we encoded a post process module to help our model extract the intent information more effectively even if the post is in any other intent domain besides those three domains.

Although we try to make our model be flexible, we still have to deal with some challenges. The most difficulty challenge is the ambiguity of natural language. This text post is an example: *"if any one want to liquidate your own Lx motorbike then call me!"*. The intent keyword of this post is implicit. While the user need to buy an old motorbike Lx, the predicted model easily extracts the intent keyword is "liquidate". So in the scope of our work, we only focus on the posts that contain explicit intents as we described in our previous paper [7]. In addition, there are several challenges that we have to face when working in natural language processing field. They are misspelling words, improper abbreviations, and free grammar...But our model will try to go through these difficulties. Overall, the

main contributions of our work are:

- We built three specific sets of labels for three selected domains and aggregated them to build the set of ten general labels
- We collected a medium-sized collection of data from discussion forums and social network, which can be used for later researches in Vietnamese intent identification.
- We addressed the problem of intent identification using the set of general labels which is domain-independent. And then we proposed a new model to solve the problem after doing some experiments carefully. Our model achieves a promising result with the average accuracy of about 80%.

The remainder of our paper is organized into five sections. Section two reviews the previous works that related to ours. In section three, we introduce our three machine learning models that we chose to solve our problem. Section four presents our proposed model. With section five, we describe our experiments. Finally, section 6 is the conclusion.

2. Related Work

Recently years, supervised learning has shown the disadvantage with the excessive growth of online data when it requires vast amount of annotated texts to create training data. Then, semi-supervised learning, transfer learning, domain-adaptation are appropriate solutions for this problem. Z. Chen et al. (2013) [1] leveraged labeled data from other domains to train a classifier for the target domain by using domain adaptation techniques. They proposed a new transfer learning method to classify the posts into two classes: intent posts (positive class) and non-intent posts (negative class). J. Wang et al. (2015) [13] proposed a graph-based semi-supervised approach to infer intent categories for tweets into six types, namely Food & Drink, Travel, Career & Education, Goods & Services, Event & Activities and Trifle. With effective information propagation via graph regularization, only a small set of tweets with category labels is needed as the supervised information. Ngo et al. (2017) [10] proposed a new method for intention detection, which leverageds labeled data in multi-source domains to improve performance in the target domain. Specifically, they used stochastic gradient descent (SGD) to optimize the aggregation process of source and target data in a Naive Bayesian framework. The method has been shown to be more effective for intention detection on the same benchmark dataset that Chen used.

Among studies that based domain-adapta -tion approach, we find the study proposed by Xiao Ding (2015) [2] seems to be the most similar to ours. They used some specific domains to learn the consumption intention. Then they attempted to transfer the CNN mid-level sentence representation learned from one domain to another by adding an adaptation layer. They also proposed to extract intention words from sentences with consumption intentions. Intention word refers to the word that can best indicate users'needs. Our work is a little different, beside the

intent keyword, we also extract necessary information related to the intent. While they used an adaptation layer, we rely on our proposed set of general labels and a post process module to solve the problem.

3. User Intent Identifying models

As almost predict model, our proposed model has two phase. The first one is the training phase, where we train the model with one of three methods, they are CRFs, Bi-LSTM, Bi-LSTM-CRFs. Training data is the data from three domain Tourism, Real estate, Transportation. The second one is the predict phase. In this phase, we used the model that we had trained in the training phase to recognize the set of labels for each instance of the new data.

3.1. Conditional Random Fields

Conditional random fields [5] are probabilistic models has shown a great success in segmenting and labeling sequence data. Given $o = \{o_1, o_2, \ldots, o_T\}$ as input observation sequence data, CRFs identifies $s = \{s_1, s_2, \ldots, s_T\}$, which is a finite set of state associated with a set of labels $l_i (l_i \in L = \{l_1, l_2, \ldots, l_M\})$, by a probability function:

$$p_\theta(s|o) = \frac{1}{Z_\theta(o)} \exp\left(\sum_{t=1}^{T} F(s, o, t)\right) \quad (1)$$

Where $Z_\theta(o) = \sum_{s'} \exp \sum_{t=1}^{T} F(s', o, t)$ is the normalizing factor to ensure that $p_\theta(s|o)$ is a probabilistic distribution, and $F(s, o, t) = \sum_i (\lambda_i f_i(s, o, t))$ is the sum of CRFs feature f_i with the feature weight λ_i correspondingly. CRFs is trained by searching the set of weights $\theta^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*)$ to maximize the log likelihood function. When the labels make the state sequence unambiguous, the likelihood function in exponential models such as CRFs is convex, thus searching the global optimum is guaranteed. It has been shown that quasi-Newton methods, such as L-BFGS, is the most efficient for this issue. In our work, we utilized pycrfsuite (https://python-crfsuite.readthedocs.io/en/latest/), which is a fast implementation of Conditional Random Fields on Python. We chose linear chain CRFs architecture because of faster training time. Features used in our model were as following:

- N-grams feature: we used unigram, bigram and trigram to capture the context of word in the posts.
- Part-of-speech (POS) tag of word was utilized to enrich linguistics features of word, i.e. user's intent is a verb or location is a noun.
- Some of entities in our data have special forms so we used word format feature to improve the accuracy in recognizing them. For example, word contains digit tend to be a point of time or price, word is initialized by a capital character tend to be a location.
- We built a dictionary to improve the learning task and using dictionary looking-up feature for unigram, bigram and trigram.

3.2. Bidirectional Long Short-Term Memory-CRFs (Bi-LSTM-CRFs)

LSTM was developed based on recurrent neural network (RNN) architecture by Hochreiter and Schmidhuber (1997) [4] and it is known to be the most effective deep learning model in natural language processing problem. In sequence tagging task, we have to take care of both past and future input features for a given time, so we chose Bi-LSTM network to do our second experiment. With this model, we can efficiently make use of past features and future features for a specific time. Following the Bi-LSTM architecture in [6], we trained our Bi-LSTM model with the following set up:

- Because our data contains both words in formal and informal style of writing so it is very hard to use pre-trained word embeddings as input to Bi-LSTM model. Instead, we utilized the embeddings learned through our network.

- We combined both word embedding feature and char embedding feature as input to Bi-LSTM to reduce the affection of words which are not in vocabulary.

Specifically, the size of char embedding and the number of char long short-term memory unit in our model are both 25. These ones for the size of word embedding and the number of word long short-term memory unit are both 100. We also used dropout technique to reduce the overfit phenomenon. Our optimization method was Adam with learning rate, learning rate decay and clip gradients initialized by 0.001, 0.9, 5.0 respectively. These hyper-parameters would be tuned together with dropout.

3.3. Bidirectional Long Short-Term Memory - CRFs (Bi-LSTM-CRFs)

Instead of making tagging independently, a CRF layer is added at the end of the tagging processing of a Bi-LSTM model. The output of Bi-LSTM layer had been considered as the input of CRFs layer and the output of CRFs layer will be the final tags. Based on the model described in [6], we utilized Bi-LSTM-CRFs model for our problem. The initialization of this model was same as the one described in Bi-LSTM model above.

4. Building the Set of Labels

With three domains that we chose to crawl the data for training model (Tourism, Transportation, Real estate), we built three specific sets of labels. We have 15 labels for Tourism, 18 labels for Real estate, and 17 labels for Transportation domain, and they are described in detail in the table 1, the table 1 and the table 1 correspondingly.

After conducting surveys carefully all the crawled data and also some others, and especially relying on the three sets of specific labels, we decided to build a set of 10 general labels. We illustrate them in the table 1 below. Some information exits in almost sort of intent domains, such as intent, object, price…, and they are used as themselves in the set of general labels. Some others are just specific for each intent domain, for example time period in Tourism domain, acreage in Real estate domain or color in Transport domain will be aggregated to make the tag description in the set of general labels.

Table 1. The domain-independent labels

Domain-Independent Label	**Abbre-viation**	**Tourism Specific Label**	**Real Estate Specific Label**	**Transportation Specific Label**
Intent	**int**	Intent	Intent	Intent
Number of Objects	**num**	Number of Objects	Number of Objects	Number of Objects
Object	**obj**	Object	Object	Object
Location	**loc**	-Destination - Point of Departure	Location	Location
Price	**prc**	Price	Price	Price
Contact	**ctt**	Contact	Contact	Contact
Context	**ctx**	Context	Context	Context
Brand	**brd**	Brand	Brand	Brand
Description	**des**	- Description of Object - Point of Time - Time Period	- Acreage - Number of Facades - Facade Size - Number of Bedrooms - Number of Bathrooms - Facade Direction - Description of Object - Equipment - Number of Floors	-Description - State - License Plate - Color - Registration Year - Model - Origin
Other	**oth**	- Name of Accommodation - Number of People - Transport	Owner	- Owner - Registration

5. Experimetal Evaluation

5.1. Experimental Data

In our work, we used the data from online forums, social media network and other websites. Specifically, we collected data for tourism domain from two main sources: *https://www.webtretho.com/forum/f110/* *and https://dulich.vnexpress.net/.* In real estate domain, data was mostly crawled from *https://batdongsan.com.vn/.* Some Facebook public groups, such as *https://www.facebook.com/groups/xemaycuhanoi*, were used for collecting data for out last selected domain, transportation. We only used the posts which have length from 30

characters up to 800 characters in order to reduce noisy data come from advertisement posts. Overall, our built dataset contains 2100 posts for tourism domain, 1200 posts in each domain of Transportation and Real Estate. After that, we had a group of 5 students to tag the data with the labels that we had built. We carefully do the cross-check among of those student work to choose the most suitable annotation. Then, we used 60% of data to train our model, 20% of data to tune the hyper-parameters. Finally, to evaluate our model we used the remaining 20% of our collected data.

5.2. *Experimental Result*

We conducted the experiments with three techniques as we mentioned above. With each individual domain, each combination of 2 domains (Tourism vs. Real estate, Tourism vs. Transportation, Real estate vs. Transportation) and the combination of all 3 domains, we carefully do the experiment with both of the set of specific labels and the set of general labels respectively. From almost of our experiment results, it can be clearly seen that it would be better to use the set of general labels when identifying user's intent from collections of data combining from various domains. On the other hand, using the sets of specific labels will mostly outperform the set of general labels when extracting intentions of user in a specific domain. The table 2 and table 3 bellow show the best chunk-based results when we do the experiment with the set of 33 specific labels and the set of 10 general labels for the combination of 3 domain datas correspondingly. This is the result when we applied CRFs method into our model.

In addition, we present the results of average accuracy when conduct experiments using three models CRFs, Bi-LSTM, Bi-LSTM-CRFs with the set of specific labels and the set of general labels respectively in the figure 1 and figure 2. In these experiments we used the combination of data from all three domains to train and test the model. We find that CRFs model outperforms the two other models. One possible reason is the size of the data for training deep learning model is too small.

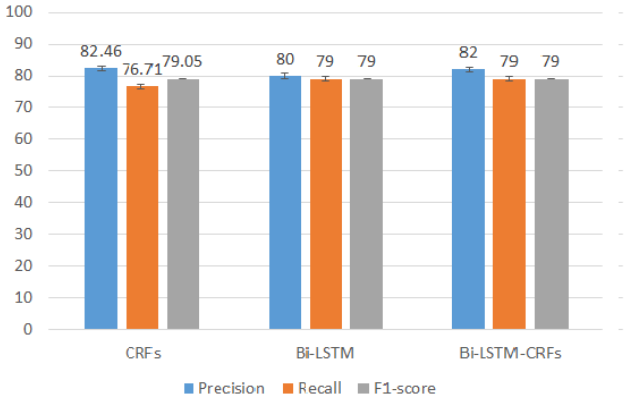

Figure 1. The Average Accuracy with Specific Labels

6. Conclusion

In this work, we present a novel method to deal with the problem of intent parsing and extraction. We call it the domain-independent intent extraction model. In this model, we propose a set of 10 general labels that is generated mainly base on three domains *Tourism, Transportation, Real Estate* and some other domain data as well. We carefully conduct more than 40 experiments to verify our assumption that the set of general labels

Table 2. The best chunk-based result with the set of specific labels

Specific Label (33)	**Precision**	**Recall**	**F1-score**
Location	71.92	75.21	73.53
Time Period	87.23	88.49	87.86
Price	85.40	85.40	85.40
Transport	62.50	46.88	53.57
Color	86.79	77.97	82.14
Facade Direction	80.00	88.89	84.21
Number of Bedrooms	87.10	60.00	71.05
Registration Year	86.96	60.61	71.43
Registration	84.78	81.25	82.98
Name of Accommodation	63.64	20.14	30.60
Point of Departure	78.26	62.07	69.23
Equipment	86.36	57.58	69.09
Description	83.04	65.03	72.94
Acreage	84.55	75.36	79.69
Number of Bathroom	100.00	83.33	90.91
Intent	88.33	86.50	87.41
Number of Objects	90.18	79.53	84.52
Number of Facades	86.96	86.96	86.96
Number of People	85.02	90.71	87.77
Facade Size	66.67	55.00	60.27
Number of Floors	84.62	91.67	88.00
Origin	94.81	79.35	86.39
Contact	90.53	91.98	91.25
License Plate	89.55	89.55	89.55
Context	44.78	38.46	41.38
Model	86.84	78.04	82.21
Description of Objects	57.62	34.39	43.07
State	63.64	50.00	56.00
Destination	81.44	68.69	74.52
Point of Time	92.45	89.09	90.74
Object	81.79	74.60	78.03
Owner	85.85	84.26	85.05
Brand	84.00	72.41	77.78
avg/total	**82.46**	**76.71**	**79.05**

Table 3. The best chunk-based result with the set of general labels

General Label (10)	**Precision**	**Recall**	**F1-score**
Num of Objects	90.99	79.53	84.87
Description	82.12	74.07	77.89
Contact	91.15	90.91	91.03
Price	86.25	85.40	85.83
Intent	88.72	86.01	87.34
Context	46.27	39.74	42.76
Object	80.70	76.84	78.72
Brand	89.58	74.14	81.13
Location	77.60	77.44	77.52
Other	81.57	69.80	75.23
avg/total	**82.57**	**77.80**	**80.06**

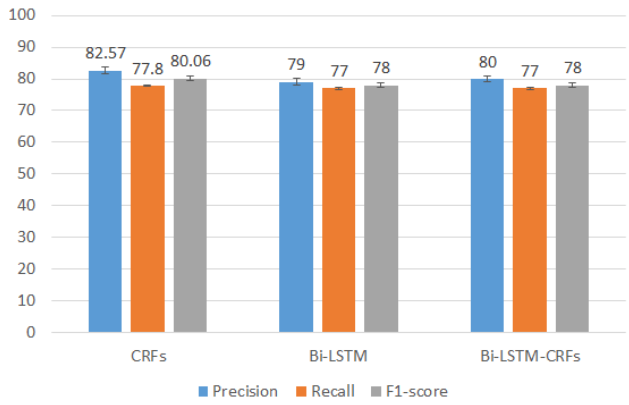

Figure 2. The Average Accuracy with General Labels

is more effective than the set of specific labels in the user intent identification task especially when intent domains are scaled up. Finally, almost of experimental results show that our proposed general labels achieve higher accuracy than specific labels in almost experiments. The average accuracies with the set of general labels are stability and almost be over 77%. Although these accuracies are not quite high, but it reconfirms that our approach is sensible. We also realize that we should improve our models and also the data to achieve higher results.

Acknowledgments

We would like to thank Dr. Tran Quoc Long, University of Engineering and Technologies, VNU Hanoi for reading the manuscript and giving us useful feedback.

References

[1] Z. Chen, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. *Identifying intention posts in discussion forums*, HLT-NAACL, 2013.

[2] X. Ding, T. Liu, J. Duan, and J.-Y. Nie. *Mining user consumption intention from social media using domain adaptive convolutional neural network*. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pages 23892395, 2015.

[3] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, S. Karwa. *Identifying purchase intent from social posts*. In Proc. of ICWSM, 2014.

[4] S. Hochreiter, and S. Jrgen. *Long short-term memory*, Neural computation pp.1735–1780, (1997).

[5] J. Lafferty, M. Andrew, and P. Fernando. *Conditional random fields: probabilistic models for segmenting and labeling sequence data*. In Proc. of ICML, 2001.

[6] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer. *Neural architectures for named entity recognition*, arXiv:1603.01360, 2016.

[7] Th.L. Luong, Th.H. Tran, Qu.T. Truong, Th.M.Ng. Truong, Th.Th. Phi, X.H. Phan. *Learning to filter user explicit intents in online vietnamese social media texts*. In Proc. of ACIID, 2016.

[8] Th.L. Luong, Qu.T. Truong, H.Tr. Dang, X.H. Phan. *Domain identification for intention posts on online social media*, In Proc. SoICT, 2016.

[9] Th.L. Luong, M.S. Cao, D.T.Le, X.H. Phan. *Intent extraction from social media texts using sequential segmentation and deep learning models*. In The 9th International Conference on Knowledge and Systems Engineering (KSE), (2017).

[10] X.B. Ngo, C.L. Le, M.Ph. Tu. *Cross-Domain Intention Detection in Discussion Forums*. In Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT), pp. 173-180, (2017).

[11] X.H. Phan, L.M. Nguyen, C.T. Nguyen. *Flexible conditional random fields*, http://flexcrfs.sourceforge.net, 2004.

[12] H. Purohit, G. Dong, V. Shalin, K. Thirunarayan, A. Shethet al. *Intent classification of short-text on social media.* IEEE International Conference on. IEEE, 2015.

[13] J. Wang, G. Cong, W.X. Zhao, X. Li. *Mining user intents in Twitter: a semi-supervised approach to inferring intent categories for tweets*. In Proc. of AAAI, 2015.