Journal of Combinatorial Optimization (2019) 37:1237–1248
https://doi.org/10.1007/s10878-018-0350-2

Speech corpora subset selection based on time-continuous utterances features

Luobing Dong[1] · Qiumin Guo[2] · Weili Wu[3]

Published online: 21 September 2018

Abstract
An extremely large corpus with rich acoustic properties is very useful for training new speech recognition and semantic analysis models. However, it also brings some troubles, because the complexity of the acoustic model training usually depends on the size of the corpora. In this paper, we propose a corpora subset selection method considering data contributions from time-continuous utterances and multi-label constraints that are not limited to single-scale metrics. Our goal is to extract a sufficiently rich subset from large corpora under certain meaningful constraints. In addition, taking into account the uniform coverage of the target subset and its internal property, we design a constrained subset selection algorithm. Specifically, a fast subset selection algorithm is designed by introducing n-grams models. Experiments are implemented based on very large real speech corpora database and validate the effectiveness of our method.

Keywords Speech corpora · Subset selection · Time-continuous utterances

1 Introduction

Artificial Intelligence (AI) based on speech recognition and semantic analysis has been gaining increasing popularity both in the research community and in industry.

✉ Luobing Dong
lbdong@xidian.edu.cn

Qiumin Guo
qmguo@mail.buct.edu.cn

Weili Wu
weiliwu@utdallas.edu

[1] School of Computer Science and Technology, Xidian University, No. 2 South Taibai Road, Xi'an 710071, Shanxi, China

[2] School of Science, Beijing University of Chemical Technology, Beijing, China

[3] Department of Computer Science, University of Texas at Dallas, Dallas, TX, USA

AI systems have been proposed for a wide range of applications such as smart manufacturing, marketing, medical care, home services, etc. A fundamental problem that underpins the effective coordinate operation of these systems is Large Corpora (Glavas and Ponzetto 2017; Matthew 2018). Many researchers suggested that the development of very large training corpora be very important, even more important than improving algorithms based on existing smaller training corpora (Clarke 2002; Schwenk and Gauvain 2005). Large corpora are ubiquitous in today's world, which contain more than a billion words each (Liu 2017). But unfortunately, simply relying upon large corpora is not sufficient (Curran and Osborne 2002). Moreover, the size of the corpora itself becomes a new serious issue for novel researches (Lin and Bilmes 2011).

Very large corpora have amounts of rich acoustic characteristics, semantic information, tokens, etc., which are very helpful for new speech recognition and semantic analysis models training. However, it also brings some trouble to the new model training, because the complexity of acoustic model training usually depends on the size of the corpora (vocabulary size, duration or storage size, etc.) (Liu 2017). Therefore, the selection of subsets with limited size and sufficient data (acoustic, semantic, etc.) is very important for corpora based researches. Most of the previous works studied the problem of reducing the influence of the size of corpora through finding subsets of large corpora based on the characteristics of their objective (Banko and Brill 2001; Boleda 2006; Braunschweiler and Buchholz 2011; Drouin 2004; Ogren 2006; Peris et al. 2017; Richmond et al. 2011). In contrast, the research of general method for subsets selection has gained less attention, but still there have been consistent efforts on designing algorithms for effectively finding parts of large corpora. McDonald et al. (1999) introduced a method of extracting parts of objects from the original corpora by ranking all the words. King et al. (2005) used greedy algorithm to create a subset of Switchboard, but the algorithm performs poorly because of the unsubmodularity of the objective function. Lin and Bilmes (2011) solved this problem. It studied the optimal selection of limited vocabulary speech corpora by presenting it as a combinatorial optimization problem on bipartite graph. These works introduced the subset selection problem and proposed some algorithms to solve it. However, there are some deficiencies in those works. First, in the design of the strategies, they failed to take into account the data among time-continuous utterances (**TCUs**) in the original corpora, which will drop much semantic information. Second, most of them simply measure the data of the resulting subset based on the number of vocabularies. This is not convincing.

As the development of technologies, more and more very large speech corpora have been created (Table 1 gives some examples of speech corpora). They usually consist of giant number of speeches from real conversations, phones, news, etc. Utterance is the unit of corpora, with many types of tags. These very large corpora promoted the research of speech recognition. Microsoft's speech and dialog research group announced 5.1% word error rate with their transcription system in 2017, which is the same to professional human transcribers'. However, the study of semantic analysis methods still falls far short of this level. Therefore, it is very important for future researches to preserve as much semantic information as possible during subset selection. Obviously, if we consider utterances separately, there will be a lot of semantic information lost because the semantics of speech are always contained in the context. For example, Table 2 shows two conversation segments from the Ubuntu Corpus,

Table 1 Some examples of Speech Corpora

Speech Corpus	Type of data	Tokens (words)	Types
HKUST	Mandarin telephone speech	1,001,895	27,210
KMITL	Reports from ACM	5,141,456 (10,000,000)	87,421
Switchboard	News, phone conversations between strangers on an assigned topic	2,400,000 (5000,000)	20,000

Table 2 Example from ubuntu corpora

No.	From	To	Utterance	Conversation
1	Dell		Well, can I move the drives?	c_1
2	Cucho	Dell	Ah not like that	c_1
3	Dell	Cucho	I guess I could just get an enclosure and copy via USB	c_1
4	Cucho	Dell	I would advise you to get the disk	c_1
5	Dell		Well, can I move the drives?	c_2
6	RC	Dell	You can't move the drives, definitely not. This is the problem with RAID :)	c_2
7	Dell	RC	Haha yeah	c_2

consisting of two sets of time-continuous utterances. Although Dell said the same sentence:" well, can I move the drives", they have different meanings.

In this paper, we investigate the problem of subset selection from very large speech corpora considering the semantics among **TCUs**. The first contribution of this paper is incorporating data contributions from **TCUs** and multi-tag constraints that are not limited to single-scale metrics (number of vocabularies) into the system model. To our best knowledge, there has been no similar research before. We analyze the submodularity of the objective function and the internal property of the target subset. Taking into account the uniform coverage of the target subset, we construct a constrained subset selection algorithm. Finally, inspired by the n-grams model (Gómez-Adorno 2018) and the experiments from **IBM** (Brown et al. 1992), we design a fast subset selection algorithm.

The remainder of this paper is organized as follows. In Sect. 2, we introduce our formulation of corpora subset selection problem and prove it is NP-hard. In Sect. 3, we prove the unsubmodularity of the objective function and some basic theorems about the internal property. In Sect. 4, we design a new subset selection algorithm which can optimize objective function, a fast algorithm based on n-grams model is presented too. In Sect. 5, we perform experiments on Switchboard-I corpus and verified our algorithm. Finally, Sect. 6 concludes the paper.

2 Problem formulation

This section presents the problem model and formalizes the language and variables used throughout this paper. Speech corpora can be divided into two types: Read speech and Spontaneous Speech (Richey 2007). The first one includes: Excerpts from books, News broadcasts, Word lists, etc. The latter includes: Conversations, Narratives, Maptasks, etc. For the simplicity of discussion, this article only focuses on the conversation subclass in the latter type. All conclusions can be easily used to other types. Given a corpora of N utterances, the goal of the subset selection algorithm is to find a subset with maximum data while satisfying size limit. A pair of utterances are said to be time-continuous if they appear in the same conversation and one right follows the other. Let U represent the original utterance set with size N. Each utterance $u_i \in U$ belongs to a conversation of the conversation set C with size m, each conversation C_j has l_j number of utterances. Let $L := (l_1, l_2, \ldots, l_m) \in R^m$ represent the size vector of C, and $X_{u_i} := (x_i^1, x_i^2, \ldots, x_i^l)$ represents the attributes or tags vector of utterance u_i. The attributes could be duration, time, tone, semantics, number of types and number of tokens, etc. For $S \subseteq U$,let $X_S = \{X_{u_i} | u \in S\}$.

Definition 1 (*Data Maximized Subset Selection Problem: DMSS*) Given characteristics and subset coverage rate limitations, **DMSS** problem can be written as the following mixed-integer (possibly nonlinear) program:

$$\begin{aligned}&\textbf{DMSS} \quad \arg\max \Omega(S) = G(S) + H(S)\\ &\text{s.t.} \quad f(X_S) \prec a\\ &|S_i|/|U_i| \prec \tau\end{aligned} \tag{1}$$

where $S \subset U$ is an utterance subset of the original corpora U, which has maximum data with the constraints. $\Omega : 2^U \to R$ is a set function, which measures the data of an utterance set. It consists of two parts: $\Omega(S) = G(S) + H(S)$. $G(S)$ measures the data in S when considering each utterance separately, and $G(S) = \sum_{u_i \in S} g(u_i)$, where $g : u_i \to R$ represents the data contained in one utterance u_i. $H(S) = \sum_{i=2}^{c_{\max}(S)} h_i(S)$ represents data generated from all **TCUs** in S, where $h_i(S)$ is the data contained in all **TCUs** of length i in S, $c_{\max}(S)$ is the max length of **TCUs** in S. $f(X_S) = \sum_{u_i \in S} f(X_{u_i})$ is a constraint function on some attributes of all utterances in S. In order to keep the result set from being concentrated in a few conversations, we set a second constraint: τ, which is the maximum coverage rate of the result set. Here $S_i = \{u_i | u_i \in S, u_i \in C_i\}$, $U_i = \{u_i | u_i \in C_i\}$.

For the simplicity of discussion, we also use a directed graph $G(U, E)$ represent the original corpora, the node set U represents the original utterance set. The edge set E represents the time-continuous relationship between utterances, if a pair of utterances are time-continuous, there is a directed edge (u_i, u_j) from u_i to u_j. Obviously, the graph consists of m paths and $\sum_{i=1}^{m} (l_i - 1)$ edges. Assume that the original corpus

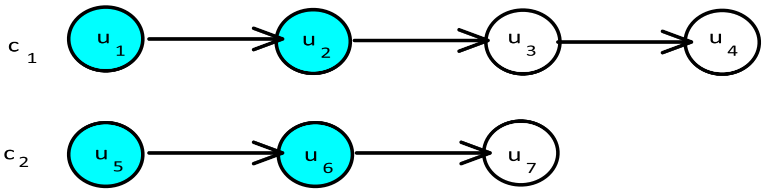

Fig. 1 The graph corresponding to Table 2. $g(u_1) = 2$, $g(u_2) = 1$, $g(u_3) = 4$, $g(u_4) = 2$, $g(u_5) = 2$, $g(u_6) = 5$, $g(u_7) = 1$, $h_2 = (\{u_1, u_2\}) = 1$, $h_2 = (\{u_2, u_3\}) = 2$, $h_2 = (\{u_3, u_4\}) = 0.5$, $h_2 = (\{u_5, u_6\}) = 3$, $h_2 = (\{u_6, u_7\}) = 0.5$, $h_3 = (\{u_1, u_2, u_3\}) = 1$, $h_3 = (\{u_2, u_3, u_4\}) = 1$, $h_3 = (\{u_5, u_6, u_7\}) = 1$, $h_4 = (\{u_1, u_2, u_3, u_4\}) = 0.5$, $f(X_{u_1}) = 8$, $f(X_{u_2}) = 5$, $f(X_{u_3}) = 15$, $f(X_{u_4}) = 8$, $f(X_{u_5}) = 8$, $f(X_{u_6}) = 18$, $f(X_{u_7}) = 2$, $f(X_S) \prec 25$, $\tau = 0.6$. Blue nodes constitute the final result set

contains only the two conversations shown in Table 2, then Fig. 1 shows the corresponding graph of the corpora. We just consider duration attribute of each utterance, and $f(X_{u_i})$ represents the duration of u_i. Utterances and conversations are numbered in the order shown in Table 2.

We aim to provide the optimal subset S, given characteristics vector X_S and subset coverage rate τ, such that the data contained is maximized. The key challenge in solving the above optimization problem **DMSS** in (1) is in unsubmodularity and NP-hardness.

Theorem 1 *The optimization DMSS problem* (1) *is NP-hard.*

Proof Consider an instance of the NP-complete Knapsack problem. Given a set of commodity $C = (c_1, c_2, \ldots, c_n)$ with weights $(a_1, a_2, \ldots, a_n)$ and profits $(p_1, p_2, \ldots, p_n)$, finding a subset of commodity whose total profit is as large as possible, and the total weight is at most b. We show that this can be viewed as a special case of **DMSS**. Given an arbitrary instance of the Knapsack problem, we define a corresponding speech corpora with n utterances: there is an utterance u_i corresponding to each commodity c_i, and it contributes $\Omega(\{u_i\})$ data which is equal to p_i. In addition, there is a character constraint value $f(X_{u_i})$ of each utterance corresponding to a_i. The Knapsack problem is equivalent to find an utterances subset S containing maximum data in this corpus with constraints $a = b$, when we do not consider the data contribution from **TCUs** and take $\tau \succ 1$. If any subset S can be obtained, then the Knapsack problem must be solvable. □

3 Solution approach

In this section, we firstly analyze the unsumodularity of **DMSS** and properties of the solution S, which could make the problem complex. Then we analyze the search space of **DMSS**, and find some important properties of it. Finally, through these properties, we obtain a solution approach to the problem.

3.1 Unsubmodularity of DMSS

Submodularity gives rise to polyhedra with very nice properties (Fujishige 2005). Researchers issue various programming problems [especially the corpus subset selection (King et al. 2005; Lin and Bilmes 2011)] which have submodular properties, such that the appropriate version of the greedy algorithm can be used to solve these problems. A set function $w : 2^V \rightarrow R$ is submodular if it satisties the following property:

$$w(A) + w(B) \geq w(A \cup B) + w(A \cap B) \tag{2}$$

for all set A and B. This means the value of considering two sets A and B separately is at least as high as the value from considering them together. If the objective function is submodularity, we can use greedy algorithm to get approximate solution in polynomial time. Unfortunately, when we consider the data contribution from time-continuous utterances in **DMSS**, it is not submodular.

Theorem 2 *The optimization DMSS problem* (1) *is not submodular.*

Proof For the objective function in (1), $\Omega(A \cup B) = G(A \cup B) + H(A \cup B)$, obviously, $G(A \cup B) = G(A) + G(B)$ and $H(A \cup B) \geq H(A) + H(B)$, then $\Omega(A) + \Omega(B) \leq \Omega(A \cup B) + \Omega(A \cap B)$, which contradicts (2), so Ω is unsubmodular. □

3.2 Solution space of DMSS

As can be seen from Theorem 2, the objective function is not submodular, so we can not simply use greedy algorithm to solve it. Therefore, analyzing the property of the solution helps us find an effective algorithm.

Consideration of **TCUs** feature makes the selected subset contain richer data information. However, it also brings trouble to the solution of the problem. For example, in Fig. 1, when we decide the data from u_2, we must consider the data from $\{u_2\}$, $\{u_1, u_2\}$, $\{u_2, u_3\}$, $\{u_1, u_2, u_3\}$, $\{u_2, u_3, u_4\}$, and $\{u_1, u_2, u_3, u_4\}$. The amount of data that related to u_2 is 6.5, (» 1 that is from only u_2). In fact, there are at least $\sum_{i=1}^{m} (l_i - 2)l_i/2$ **TCUs** which generate more data. This is far more than the data generated without considering time continuousness, which makes the problem more complex.

TCUs feature also bring another trouble: it makes the search space of the target subset larger. In the previous example, because of the consideration of all **TCUs** containing u_2, the semantic data can be reserved as much as possible during the subset selection process, but it also causes the exponential growth of the search space of the problem. Fortunately, **TCUs** in the result subset show some patterns like Theorem 3.

Theorem 3 *For* $S^* = \arg\max \Omega(S)$, $S \subseteq C_i$, $|S|= k$, *if* $T \subseteq S^*$ *consists of* $l-TCUs$ *and* T *is not adjacent to any utterance in* $S^* - T$, *then* $\Omega(T)$ *must be among the largest* $l(k-l)+1$ *elements of* $\{\Omega(X)|\ X\ are\ l - TCUs\ of\ C_i\}$.

Proof When $l = k$, all utterances in S^* are time-continuous, and $T = S^*$, $l(k-l)+1 = 1$. The theorem holds.

For any $l \prec k$, assume the conclusion is not true. Let $X_l = \{X | X\ are\ l-TCUs\ of\ C_i\}$. $\forall u \in S^* - T$, there are at most l elements in X_l which contain u. Therefore there are at most $l(k-l)$ elements in X_l that contain utterance of $S^* - T$. So, $\exists T^{'} \in X_l, T^{'} \cap (S^* - T) = \emptyset$ and $\Omega(T^{'}) \succ \Omega(T)$. Let $S^{'} = (S^* - T) \cup T^{'}$, then $\Omega(S^{'}) \succ \Omega(S)$, which is a contradiction. □

In the proof of Theorem 2, we can find that $G(A) + G(B) = G(A \cup B) + G(A \cap B)$, but $H(A) + H(B) \leq H(A \cup B) + H(A \cap B)$. Therefore, the introduction of the **TCUs** feature is the cause of the unsubmodularity. At the same time, we also realize that, if $A \cap B = \emptyset$ and there are other utterances between A and B (especially when A and B belong to different conversations), $H(A) + H(B) = H(A \cup B) + H(A \cap B)$. Interestingly, this also helps simplify the objective issue, we will see it in Sect. 4.

3.3 Solution approach

We know the size of speech corpora is usually very big, and the amount of time and resources needed to solve this problem grows exponentially with it. Fortunately, the single conversation is much smaller, moreover, the utterances in different conversations are not time continuous. Therefore, we can find the subsets from each conversation and then merge all the subsets. However, Theorem 1 tell us that, even for one single conversation, the **DMSS** problem is NP-hard, so heuristics are needed to find a good solution faster.

From the previous section, we know that the time continuousness of utterance makes the problem space very large. Researchers usually studied the semantic effects of continuous words in sentences through n-grams models. IBM researchers found through experiments that in the n-grams models, when n is equal to 3, the generated phrases are almost impossible to appear in the actual language, let alone more than 3 (Brown et al. 1992). Therefore, most of the semantic researches based on n-grams models only consider 2 or 3-grams (Gómez-Adorno 2018; Kumar and Satyanarayana 2017; Walter et al. 2017), ignoring the larger ones. Since it is so for a single word, it should be more like this for utterances in conversations. Therefore, we do not need to consider the data contribution of **TCUs** longer than 2. Moreover, from Theorem 3, we know that we only need to search from $z_1 = |U_i|\tau$ unit utterances and $z_2 = 2|U_i|\tau - 3$ 2-**TCUs** (**TCUs** of 2 length), which are much lesser than $2|U_i|$. This will allow us to get a solution faster. At the same time, for a single conversation, we can still use greedy algorithms in our algorithm, because the search space will be very small after preprocessing (we can see this in Sect. 4).

4 2-grams based algorithm

In this section, we propose one 2-grams based algorithm (shown in Algorithm 1) to solve the **DMSS** problem. The idea is to reduce search space by selecting part of the largest 1-**TCUs** and 2-**TCUs**. Then the selected **TCUs** are merged based on their

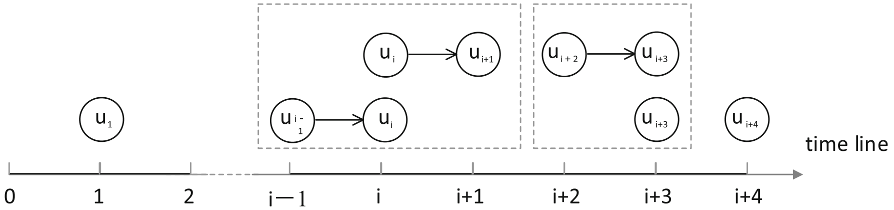

Fig. 2 A merging example. The abscissa represents the time axis. Nodes in the same dashed box overlap in occurrence time and need to be merged. After merging, we get the final utterances set:$\{u_1, \{u_{i-1}, u_i, u_{i+1}\}, \{u_{i+2}, u_{i+3}\}, u_{i+4}\}$

occurrence time. Finally, the subset can be selected by greedy algorithm. Thus, the partitioning consists of three phases.

4.1 Initialization and search space creation

Line 1 to Line 8. We make the necessary variable initialization in this section. And, more importantly, the search space is reduced by sorting and filtering. Line 4 means that the amount of data contained in each utterance in U_i (U_i is the set of utterances in conversation C_i) is sorted and the largest $l_i\tau$ utterances are placed in U_i^1. Line 5 to Line 6 select all 2-**TCUs** and place them in $U_i^{'}$. Line 8 is similar to Line 4, but it should be noted that $u \in U_i^{'}$ is a 2-**TCUs**, so its data contains not only the data of each utterance, but also the compound data from them two together. The search space of the problem will be significantly reduced after the filtering of Line 4 and Line 8.

4.2 Merging

Line 9, and the function detail is shown in Algorithm 2. We merge utterances overlapping in occurrence time to further reduce search space. Figure 2 shows a merge example. It should be noted that in order to avoid excessive very long **TCUs**, which will make the result subset only in a few segments of the original conversation, we do not merge time-continuous **TCUs**. There can be many kinds of merging strategies. For example, we can merge all the continuous utterances one by one. If merging one utterance fragment in makes the sum of the constrained attribute value be greater than the predetermined value y_i, no merging will be done. We can also merge as many fragments as possible. If the constraint attribute value of the result segment is greater than y_i, we can perform a split such as dichotomy to ensure that the value of the constraint attribute of the obtained result utterance segment does not exceed the limit. Due to space limitations, we only give the first method here (Algorithm 2).

4.3 Subset selection

Line 10–Line 24. After the second phase, although some of the remaining utterance segments could also be time continuous, and the objective function in formula (1) is still unsubmodular, we can use greedy algorithm to find the best subsets by choosing better initial nodes.

Algorithm 1 **DMSS** Algorithm

```
Input $U$ : Original Utterances Set; $C$ :Original Conversations Set; $Y$ :Original characteristics
    value of $U$ $a$ : Maximum Value of Total Constraint Characteristics; $\tau$ : Maximum Coverage
    Constraint.
OutPut: $S$ : Maximum Subset of $U$
$S \leftarrow \varnothing, m = |C|, n = |U|$
for $i = 1 \rightarrow m$ do
    $S_i \leftarrow \varnothing, l_i \leftarrow |U_i|, y_i \leftarrow l_i a / m, U_i^{'} \leftarrow \varnothing$
    $U_i^1 \leftarrow \{Sort(U_i, \Omega(u))\}(1 : l_i \tau)$ for all $u \in U_i$,
    for $j = 2 \rightarrow l_i$ do
        $U_i^{'}(j) \leftarrow \{U_i(j), U_i(j+1)\}$
    end for
    $U_i^2 \leftarrow U_i^2 + \{Sort(U_i^{'}, \Omega(u))\}(1 : 2l_i \tau - 3)$ for all $u \in U_i^{'}$
    $U_i^3 = Merge(U_i^1, U_i^2)$
    $S_i = S_i + FindL \arg est(U_i^3)$
    $z = 1y^{'} = 0$
    while $z \prec |U_i^3|$and $y^{'} \prec y_i$ do
        for all $u$ such that $u \in U_i^3$ and $u \notin S_i$ do
            datraincre = the data increased by adding $u$ in $U_i^3$ to $S_i$
        end for;
        $newu = \arg\max_u(dataincre)$
        if $y^{'} + Y(newu) \leq Y_i$ and $Length(newu) + |S_i| \leq \tau l_i$ then
            $S_i = S_i + newu$
            $z = z + 1$, $y^{'} = y^{'} + Y(newu)$
        else break
        end if
    end while
    $S \leftarrow S + S_i$
end for
```

5 Experimental results

We performed experiments on Switchboard-I corpus and verified our algorithm. In our experiments, we measured the amount of data contained in each utterance by its correlation coefficient to the topic of the conversation. Utterance tags were used to measure the amount of compound data produced by a 2-**TCUs**. For example, if the tag of an utterance is “aa”, “qy”, “qw”, “bk”, “nn”, etc., the amount of compound data from the utterance and its adjacent utterances will be bigger. In addition, for the characteristics constraint, we used the total duration.

Algorithm 2 Merge Algorithm

```
function MERGE($U_i^1$, $U_i^2$, $y_i$)
    $U_i^3 \leftarrow \varnothing$
    for all $u \in U_i^2 \cup U_i^1$ do
        if $\exists u' \in U_i^3$ then
            $u'' = u'$ merge $u$
            if $f(u') \prec y_i - f(u'')$ then
                $U_i^3 = U_i^3 - u'$
                $U_i^3 = U_i^3 + u''$
            else
                $U_i^3 = U_i^3 + u$
            end if
        else
            $U_i^3 = U_i^3 + u$
        end if
    end for;
    return $U_i^3$
end function
```

Because the test data set and result set are too large, we can't list all of them here. We put the lgorithm source code and result set at the following URL: https://github.com/18700197078/Ed/tree/master/src. Table 3 shows some of the result from our experiments. Each conversation was divided into several time-continuous segments that represent the semantics of the dialogue. While significantly reducing the size of the corpora, semantic information was remained as much as possible, which was consistent with our expectations. The subset selection process was very fast. For example, for conversations with sequence number 4327, regardless of τ being 0.1, 0.2, or 0.3, the result set is always concentrated in two specific segments (23 ~ 25, 72 ~ 77). These two segments exactly represent the semantics of the entire dialogue. The average search time for each conversation was 5 ms, and the search process of the entire corpora could be completed within 5 min. Although this is only the speed under the simple constraints and measurement method of data as we set, it is also sufficient to show that our algorithm is relatively high in efficiency. This would provide a very good basis for the selection of the ideal prospective subset, and it could also effectively shorten the development period of the AI algorithm.

Table 3 Experiments results

τ	Con no.	Con size	Result	Subset size	Time (S)
0.1	4327	88	23 ~ 25, 72 ~ 77,	9	125
	4330	72	35 ~ 37, 43 ~ 48, 88 ~ 89	11	
	4171	108	32 ~ 39, 52 ~ 53,	10	
	4321	110	23 ~ 30	8	
	4329	137	30 ~ 33, 72 ~ 77, 137 ~ 139	13	
0.2	4327	88	22 ~ 27, 72 ~ 83	18	151
	4330	72	2 ~ 5, 35 ~ 38, 43 ~ 49, 58 ~ 59, 88 ~ 91	11	
	4171	108	32 ~ 40, 43 ~ 53, 107 ~ 108	22	
	4321	110	20 ~ 33	14	
	4329	137	19 ~ 20, 30 ~ 37, 63 ~ 79,	27	
0.3	4327	88	19 ~ 29, 72 ~ 79, 101 ~ 107	26	204
	4330	72	33 ~ 38, 40 ~ 50, 83 ~ 101,	32	
	4171	108	32 ~ 63	32	
	4321	110	6 ~ 63	32	
	4329	137	30 ~ 37, 63 ~ 87, 137 ~ 143	41	

Con no., conversation no.; con size, conversation size measured by utterances number; result, result subset represented by its members' utterance no.; subset size, measured by utterances number of result subset; time, time used for subset selection

6 Conclusion

The speech corpora subset selection problem considering time continuous utterances' compound data contribution brings new challenges, and is not addressed by the previous researches. In this paper, we propose a search space reducing process and 2-grams based subset selection algorithm. We first analyze the unsubmodularity, NP-hardness, and properties of the problem. With the estimated range of the result set, a **TCUs** filtering and merging process is proposed to get the smaller search space. Finally, greedy algorithm can be used by appropriately selection of initial **TCUs**, and its speed is fast. Experiments on Switchboard-I prove that the algorithm is efficient and correct.

Acknowledgements This work is partly supported by National Science Foundation under Grant 1747818, and the Fundamental Research Funds for Central Universities (JB161004).

References

Banko M, Brill E (2001) Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting on association for computational linguistics—ACL'01. Toulouse, France, pp 26–33

Boleda G et al (2006) CUCWeb: a Catalan corpus built from the Web. In: Wac'06 processing of the 2nd international workshop on web as corpus. April. Trento, Italy, pp 19–26

Braunschweiler N, Buchholz S (2011) Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In: Proceedings of the annual conference of the international speech communication association, Interspeech. August. Florence, Italy, pp 1821–1824

Brown PF et al (1992) Class-based n-gram models of natural language. Comput Linguist 4(18):467–479

Clarke CLA et al (2002) The impact of corpus size on question answering performance. In: Proceedings of the 25th annual international ACM SIGIR conference on research development on information retrieval, pp 369–370

Curran JR, Osborne M (2002) A very very large corpus doesn't always yield reliable estimates. In: Proceedings of the 6th conference on natural language learning—COLING-02. Vol. 20. Stroudsburg, PA, USA, pp 1–6

Drouin P (2004) Detection of domain specific terminology using corpora comparison. In: Proceedings of the 4th international conference on language resources and evaluation. Lisbon, Portugal, pp 79–82

Fujishige S (2005) Submodular functions and optimization, vol 58. C. Elsevier, Amsterdam, pp 315–363

Glavas G, Ponzetto SP (2017) Dual tensor model for detecting asymmetric lexico-semantic relations. In: Proceedings of the 2017 conference on empirical methods in natural language processing. September. Copenhagen, Denmark, pp 1757–1767

Gómez-Adorno H et al (2018) Document embeddings learned on various types of n-grams for cross-topic authorship attribution. In: Computing September, pp 1–16

King S, Bartels C, Bilmes J (2005) SVitchboard 1: small vocabulary tasks from switchboard 1. In: Ninth European conference on speech communication and technology. Lisbon, Portugal, pp 2–5

Kumar VV, Satyanarayana N (2017) Probability of semantic similarity and N-grams pattern learning for data classification. In: Global journal of computer science and technology, pp 1–5

Lin H, Bilmes J (2011) Optimal selection of limited vocabulary speech corpora. In: Proceedings of the annual conference of the international speech communication association, interspeech, Florence, Italy, pp 1489–1492

Liu Y et al (2017) SVitchboard II and FiSVer I: high-quality limited-complexity corpora of conversational English speech. In: Proceedings of the annual conference of the international speech communication association, interspeech, vol 42, pp 122–142

Matthew S (2018) An extensible schema for building large weakly-labeled semantic corpora. Proced Comput Sci 128:65–71

McDonald G, Macdonald C, Ounis I (1999) Finding parts in very large corpora, vol June, College Park, pp 57–64

Ogren PV et al (2006) Building and evaluating annotated corpora for medical NLP systems. In: AMIA annual symposium proceedings/AMIA symposium. AMIA symposium 36.2003, p 1050

Peris Álvaro, Chinea-Rios Mara, Casacuberta Francisco (2017) Neural networks classifier for data selection in statistical machine translation. Prague Bull Math Linguist 108(1):283–294

Richey C (2007) https://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf. Accessed 6 Feb 2007

Richmond K, Hoole P, King S (2011) Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In: Proceedings of the annual conference of the international speech communication association, interspeech. August. Florence, Italy, pp 1505–1508

Schwenk H, Gauvain J-L (2005) Training neural network language models on very large corpora. In: Proceedings of the conference on human language technology and empirical methods in natural language processing—HLT'05. Vancouver, B.C., Canada, pp 201–208

Walter L, Radauer A, Moehrle MG (2017) The beauty of brimstone butterfly: novelty of 290 patents identified by near environment analysis based on text mining. Scientometrics 111(1):103–115