# Supplementary Materials for Phrase-based Compressive Cross-Language Summarization

Jin-ge Yao Xiaojun Wan Jianguo Xiao
Institute of Computer Science and Technology, Peking University
{yaojinge, wanxiaojun, xiaojianguo}@pku.edu.cn

## 1 Proof for Theorem 1

### 1.1 Background

Here we briefly describe some background knowledge of submodular function maximization. Formally, submodularity is defined as a property of set functions for a finite discrete universe set $U$. A set function $F : 2^U \rightarrow \mathbb{R}$ is said to be *submodular* if: $\forall S, T \subseteq U$, we have $F(S) + F(T) \geq F(S \cup T) + F(S \cap T)$. An equivalent but more intuitive description is known as the *diminishing returns* property: $\forall S \subseteq T \subseteq U \setminus u$, a submodular function $F$ must satisfy $F(S \cup \{u\}) - F(S) \geq F(T \cup \{u\}) - F(T)$. That is to say, the addition of $u$ will bring more benefit for a smaller set $S$ compared with a larger set $T \supseteq S$. The property of diminishing returns is suitable for document summarization purposes. Once we decide to add a sentence in the summary, the gain of information should be small when there already exists sufficient information in the summary.

A set function $F$ is *monotone nondecreasing* if $\forall S \subseteq T, F(S) \leq F(T)$. Monotone nondecreasing submodular functions will simply be referred to as *monotone submodular* for short.

Extractive summarization can be easily modeled as submodular maximization problems when the summary scoring function is defined as a submodular function of sentences.

Normally, the budgeted submodular maximization problem can be efficiently solved by a greedy algorithm (similar to Algorithm 1 in the submitted paper) with provable guarantees of a constant approximation factor to the optimal solution.

### 1.2 Proofs

Here we give a formal proof for Theorem 1 in our main submission. This part is inspired by [1] which provides an approximation guarantee for a fully submodular scoring function.

Notation: $S_i$ denotes the collection of selected items (compressed sentence) in the $i$-th iteration, with the item added in this iteration denoted as $s_i$. $C(S)$ denotes the total cost of the collection $S$ (similarly $C(s)$ denotes the cost of a single item $s$) and $B$ denotes the budget. We assume the optimal solution is $OPT$.

We use as a shorthand notation for the distortion term: $\Delta(s_i) = dist(y(s_i)) \geq 0$. We also generalize this notation to sets: $\Delta(S) = \sum_{s \in S} dist(y(s))$. This term decomposes with every sentence $s$ inside. We also denote the monotone submodular part of $F$ as $f$, therefore we have $F(S) = f(S) + \eta\Delta(S)$. For the $i$-th iteration, denote $OPT \setminus S_{i-1} = \{u_1, \ldots, u_m\} \equiv u_1^m$.

As there is often a distortion limit $|start(p) - 1 - end(p)| \leq \delta$ in phrase-based MT systems for ensuring decoding efficiency and translation quality, we introduce a reasonable assumption [1]:

**Assumption 1.**

$$\Delta(OPT \cup S_{i-1}) - \Delta(OPT) \leq \gamma,$$

*or equivalently,*

$$\Delta(S_{i-1}) + \Delta(OPT \setminus S_{i-1}) - \Delta(OPT) \leq \gamma,$$

where $\gamma$ is a bounded constant that is relatively small in practice compared with the value of the submodular score $f(S)$. $\eta < 0$ is the common distortion parameter.

**Lemma 2.** *For $i = 1, \ldots, l+1$, we have*

$$F(S_i) - F(S_{i-1}) \geq \frac{C(s_i)}{B}(F(OPT) - F(S_{i-1})) + \frac{C(s_i)\eta\gamma}{B}.$$

*Proof.*

$$\begin{aligned} F(OPT) - F(S_{i-1}) &= f(OPT) + \eta\Delta(OPT) - f(S_{i-1}) - \eta\Delta(S_{i-1}) \\ &\leq f(OPT) - f(S_{i-1}) + \eta\Delta(OPT \setminus S_{i-1}) - \eta\gamma \\ &\leq f(OPT \cup S_{i-1}) - f(S_{i-1}) + \eta\Delta(OPT \setminus S_{i-1}) - \eta\gamma \\ &= F(OPT \cup S_{i-1}) - F(S_{i-1}) - \eta\gamma. \end{aligned}$$

using Assumption 1 in the first and monotonicity of $f$ in the second inequality.

Denote $Z_j = F(S_{i-1} \cup u_1^j) - F(S_{i-1} \cup u_1^{j-1})$, we have

$$\frac{Z_j}{C(u_j)} \leq \frac{f(S_{i-1} \cup u_j) - f(S_{i-1}) - \Delta(u_j)}{C(u_j)} = \frac{F(S_{i-1} \cup u_j) - F(S_{i-1})}{C(u_j)} \leq \frac{F(S_i) - F(S_{i-1})}{C(s_i)}$$

using submodularity of $f$ in the first and the greedy selection strategy (Algorithm 1 in the main submission) in the second inequality. Since $\sum_{j=1}^m C(u_j) \leq B$, it holds that

$$F(OPT) - F(S_{i-1}) \leq \sum_{j=1}^m Z_j - \eta\gamma \leq B\frac{F(S_i) - F(S_{i-1})}{C(S_i)} - \eta\gamma.$$

□

[1] Note that this assumption will not directly lead to the main result since we are optimizing a different objective function and using a different greedy selection strategy, compared with the original submodular maximization case presented in previous work.

**Lemma 3.** *For $i = 1, \ldots, l+1$, we have*

$$F(S_i) \geq [1 - \prod_{k=1}^{i}(1 - \frac{C(S_k)}{B})]F(OPT) + \eta\gamma.$$

*Proof.* Note that the budget constraint will make $C(S_i) \leq B$ for any $i$. We give a proof by induction. For $i = 1$, we have

$$F(S_1) \geq \frac{C(S_1)}{B}F(OPT) + \frac{C(S_1)\eta\gamma}{B} \geq \frac{C(S_1)}{B}F(OPT) + \eta\gamma$$

due to Lemma 2 and $C(S_i) \leq \sum_i C(S_i) \leq B$. For $i > 1$:

$$\begin{aligned}
F(S_i) &= F(S_{i-1}) + F(S_i) - F(S_{i-1}) \\
&\geq F(S_{i-1}) + \frac{C(S_i)}{B}[F(OPT) - F(S_{i-1})] + \frac{C(S_i)\eta\gamma}{B} \\
&= (1 - \frac{C(S_i)}{B})F(S_{i-1}) + \frac{C(S_i)}{B}F(OPT) + \frac{C(S_i)\eta\gamma}{B} \\
&\geq (1 - \frac{C(S_i)}{B})[(1 - \prod_{k=1}^{i-1}(1 - \frac{C(S_k)}{B}))F(OPT) + \eta\gamma] + \frac{C(S_i)}{B}F(OPT) + \frac{C(S_i)\eta\gamma}{B} \\
&= (1 - \prod_{k=1}^{i}(1 - \frac{C(S_k)}{B}))F(OPT) + \eta\gamma.
\end{aligned}$$

using Lemma 2 in the first and the induction hypothesis (Lemma 3 is true for $i - 1$) in the second inequality. □

We restate our main result here, i.e. Theorem 1 in the main submission:

**Theorem 4.** *If Algorithm 1 outputs $S^{greedy}$, we have*

$$F(S^{greedy}) \geq \frac{1}{2}(1 - e^{-1})F(OPT) + \frac{1}{2}\eta\gamma.$$

*Proof.* Exactly the same as the proof of Theorem 1 in [1], using our Lemma 2 and Lemma 3 and changing the constant term into $\eta\gamma$. □

# 2 Example System Outputs

Figure 1 lists the summaries for the first document set (D04) in the DUC 2001 dataset, produced by systems in comparison. The summaries are produced under the sentence budget setting, i.e. limiting total number of sentences no more than five. Those who can read Chinese texts will observe that our compressive system tries to compress sentences by removing relatively unimportant phrases. The effect of translation errors can also be reduced since those incorrectly translated words will be dropped for having low information gains. In some cases the grammatical fluency can even be improved from sentence compression, as redundant parentheses may sometimes be removed.

CoRank: 但仅在佛罗里达州的损失，安德鲁飓风成为美国最昂贵的保险的灾难。过去的严重飓风美国，雨果，袭击南卡罗来纳州于1989年，耗资从保险损失行业42亿美元，但造成的总伤害的估计60亿美元和100亿美元之间不等。美国城市沿墨西哥湾的阿拉巴马州到得克萨斯州东部是在风暴手表昨晚安德鲁飓风向西横跨佛罗里达州南部席卷后，造成至少八人死亡和严重的财产损失。美国的保险公司希望在佛罗里达州的安德鲁飓风的结果支付估计73亿美元（37亿英镑）-迄今为止最昂贵灾难的行业有史以来遇到。虽然已出现在保险理赔的水平一些初步猜测，昨日的图来自美国保险服务集团，财产险保险公司行业协会的财产索赔服务部门。

PBES: 凯特女士硬朗，紧急服务在佛罗里达州的戴德县，承担了风暴的冲击主任估计，安德鲁已经造成150亿美元到200亿美元的损害（75亿英镑，100亿英镑）。雨果飓风，袭击东海岸在1989年9月，花费了保险业约42亿美元。美国城市沿墨西哥湾的阿拉巴马州到得克萨斯州东部是在风暴手表昨晚安德鲁飓风向西横跨佛罗里达州南部席卷后，造成至少八人死亡和严重的财产损失。过去的严重飓风美国，雨果，袭击南卡罗来纳州于1989年，耗资从保险损失行业42亿美元，但造成的总伤害的估计60亿美元和100亿美元之间不等。最初的报道称，至少有一人已经死亡，75人受伤，数千取得沿着路易斯安那州海岸无家可归，14证实在佛罗里达州和死亡三巴哈马群岛后。

PBCS: 紧急服务在佛罗里达州的风暴的冲击主任估计，安德鲁已经造成150亿美元到200亿美元的损害（75亿英镑，100亿英镑）。雨果飓风，袭击东海岸在1989年9月，花费了保险业约42亿美元。美国城市沿墨西哥湾的阿拉巴马州到得克萨斯州东部是在风暴昨晚安德鲁飓风向西横跨佛罗里达州南部后，造成至少八人死亡和严重的财产损失。严重飓风美国，雨果袭击南卡罗来纳州于1989年，保险损失行业42亿美元，但造成的总伤害的估计60亿美元和100亿美元之间。最初的报道称，至少有一人死亡，75人受伤，数千取得沿着路易斯安那州海岸无家可归，14在佛罗里达州和死亡巴哈马群岛后。

Human: 飓风安德鲁袭击佛罗里达州和路易斯安那州，造成17人死亡，经济损失大概150亿美元。南佛罗里达州，尤其是戴德县，遭受到最严重的破坏，每八个房屋中就有一个被毁，15万人无家可归。在路易斯安那州，安德鲁没有波及到新奥尔良，其造成的损失大部分为农村地区。保险理赔对于美国公司可达80亿美元，另外10亿美元用于英国公司。佛罗里达州的很多家庭和企业都没有投保。联邦紧急事务管理局对于安德鲁的受害者，未能提供及时或充足的联邦援助。布什总统访问了该地区，并承诺提供援助。

Figure 1: Example system outputs

# References

[1] Andreas Krause and Carlos Guestrin, A Note on the Budgeted Maximization of Submodular Functions, Technical Report, 2005.