口语对话中冗余词汇识别方法研究[1]

翟飞飞，宗成庆

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：冗余现象是口语对话中普遍存在的特殊语言现象之一，它的存在常常会影响口语句子的理解和翻译。本文基于真实口语对话语料对冗余现象进行了分析，并在词汇层面对冗余现象进行了分类，然后对口语中的冗余词汇进行了统计识别方法研究。通过对冗余词汇处理前后的口语句子翻译实验，结果表明，预先对冗余现象进行处理，能够改善口语翻译的译文质量。

关键词：冗余现象；最大熵分类器；支持向量机；条件随机场

中图分类号：TP391

Research on Recognition of Fillers in Spoken Dialogs

Feifei Zhai and Chengqing Zong

(National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China)

Abstract: Fillers and redundancy are most common phenomena in spoken dialogs. It always influences the results of spoken language understanding and translation system. Based on analysis and statistical classification of fillers in lexical level of spoken dialog corpus, we propose statistical methods to recognize the fillers. Experiments on translation of the spoken sentences before and after processing of the fillers have been conducted. The experimental results have shown that the performance of spoken language translation system is significantly improved if the fillers are processed before translating.

Key Words: Fillers; maximum entropy classifier; SVM; CRFs

1. 引言

口语是人与人之间相互交流最直接、最方便、最自然的语言使用方式。与书面语相比，由于人在说话过程中思维和交流的需要，口语句子中往往存在很多不同的语言现象，包括间断、省略、重复、修正、词序颠倒和冗余等。对于口语中的这些特殊语言现象的研究和处理，是目前口语分析所面临的一个难点[1]。

文献[1]对汉语口语中的冗余现象进行了定义，即当说话人思维停顿或不连贯的时候，往往不自觉地在句子中添加一些词汇来保持语气和句子的连续性，这些添加的词汇就构成了冗余。冗余现象是句中的多余成份，去掉这些成份不会对句子的意思和结构产生影响。例如：口语句子：“是那个如家快捷吗？”中的“那个”一词在本句中为冗余成份。

在英语口语语言现象的自动处理方面，人们对重复、修订等现象进行了大量的研究，冗余现象的

[1] 本文的研究工作得到国家自然科学基金项目的资助（资助号：60975053、61003160，60736014）、中国科学院对外合作交流项目的资助和中国新加坡数字媒体研究院的资助。

处理主要是作为上述工作的一个子任务来完成的。文献[2]使用规则的方法来识别口语句子中的冗余成份。文献[3]使用条件随机场(CRFs)，文献[4]使用最大熵模型在识别其他特殊语言现象的同时对冗余成份进行了识别。文献[5]对删除冗余成份后造成的影响进行了实验，他们发现提前删除冗余成份能够提高依存句法分析的准确率。

在汉语口语语言现象的自动处理方面，针对口语冗余现象处理的工作并不是很多。文献[6]介绍了中科院自动化所和社科院语言所联合建设的汉语口语对话语料库，它涉及了旅馆预订、机场信息咨询、以及餐馆订餐等领域，并对口语的各种特殊现象进行了标注。文献[7]对 100 多段涉及旅馆预订领域的口语对话语料进行了统计分析，冗余现象在口语的特殊语言现象中所占比例为 4.7%。文献[8]详尽的分析了真实口语语料中的冗余现象,对冗余现象涉及的词汇长度，词汇分布进行了统计，并对冗余现象出现的特征进行了详细分析并提出了一种面向中间表示的口语解析方法，得到了较好的鲁棒性和正确率，但文献[8]并没有对冗余现象进行针对性的处理。冗余现象是影响口语句法和语义分析的重要因素之一[1]。当系统对口语句子进行分析或自动翻译时，冗余现象的存在常常会降低分析准确率,影响译文的质量。因此，如果能够在口语句子分析之前把这部分冗余成份进行过滤或去除，就可以有效地减轻后续分析模块的负担。当然，这个预处理过程本身也可以看作是口语理解的一项任务。本文基于真实汉语口语语料库对冗余现象进行分析，并对汉语口语中的冗余词汇识别方法进行了研究。

本文的其余部分按如下安排：第 2 节我们对口语对话语料中的冗余现象进行分析；第 3 节介绍了本文进行冗余词汇识别和处理的思路；第 4 节简单介绍了所使用的统计方法：最大熵模型，支持向量机，以及条件随机场；第 5，6 节介绍了冗余词汇识别的方法并给出了实验结果和分析；第 7 节对冗余词汇处理前后的口语句子进行了翻译实验、并进行人工打分以及结果分析；最后对全文进行总结。

2. 口语对话语料统计结果及分析

2.1 口语语料预处理

本文使用中科院自动化所和社科院语言所联合建设的 CASIA-CASSIL 口语对话语料库的部分语料，包括：旅馆咨询领域 44 段 1522 句、机场信息问询领域 55 段 1471 句、以及餐馆订餐领域 41 段 1509 句。为了便于使用统计方法对口语中的冗余词汇进行识别，我们首先把口语句子中的冗余现象分为以下四类：

①语气词性质的冗余词汇，即在句子中表示语气的词，例：“啊”、“呀”等，记作 prt^y；

②代词性质的冗余词汇，即词性为代词，且可以作为冗余词汇的词汇，例：“这个”、“那”等，记作 prt^r；

③副词性质的冗余词汇，即词性为副词，且可以作为冗余词汇的词，例：“就”等，记作 prt^d

④插入语性质的冗余词汇，除上述三类的冗余都分入此类，例“就是说”,“就是”等，记作 prt^p；

另外，为了后续描述的方便，我们把后面三类的合并，即②③④的并集记作 prt^rdp。

我们按照上述四类对所使用的口语对话语料中的冗余词汇进行标注。在标注之前，我们首先对语料进行了如下预处理：(1)由于真实的口语转成的文本中并没有标点符号信息，而 CASIA-CASSIL 语料中含有标点符号信息，因此在标注之前，我们首先删除了语料中的标点符号。(2)由于我们没有发现专用于真实场景口语对话语料的分词工具，我们使用分词工具 ICTClass 对语料进行分词和词性标注。然后按照我们的分类模式对语料中的冗余词汇进

行标注[2]，标注的同时我们也对分词错误进行了修改。完成上述工作后，就得到了本文所使用的语料。

2.2 统计结果及分析

对语料进行预处理之后，我们对使用的口语对话语料进行了统计分析。在旅馆咨询领域、机场信息问询领域、以及餐馆订餐领域三个领域的口语语料中，冗余词汇在总词汇中所占的比例分别为11.97%、15.23%、和14.4%。这说明在真实的口语对话中，有至少11%的词汇不仅对口语的语法分析和语义理解没有帮助，而且会产生不利影响。其他的统计结果如表1。

表1. 真实口语语料统计结果

冗余类别 / 语料领域	语气词性质(%)	代词性质(%)	副词性质(%)	插入语性质(%)
旅馆咨询	56.53	21.01	15.46	7
机场信息问询	56.81	21.27	15.04	6.88
餐馆订餐	79.01	13.44	5.51	2.03

表格中的数据代表当前语料中当前类别的冗余词汇在所有冗余词汇中所占的比例。从表中可以看出，无论哪个领域，语气词性质的冗余词汇(prt^y)在冗余词汇中所占比例最大，超过50%，说明人们在口语对话中语气词的使用较多，这也符合一般的规律。前三类冗余词汇所占比例之和可以达到93%以上，这说明冗余词汇在词典中的分布主要集中在语气词、副词、代词三种词性上，这也为我们的冗余成份识别提供了便利。

3. 本文的思路

经由上面的分析，我们知道冗余词汇的分布相对集中，而冗余现象则主要是由冗余词汇构成[3]，因此本文以冗余词汇作为冗余现象识别和处理的基本单位。本文对冗余词汇进行识别和处理的任务就是识别并删除口语对话语句中的冗余词汇。删除句子中的冗余词汇后，就可以把处理好的口语句子送入后续应用模块进行后续处理，如口语句法分析，口语机器翻译等。因此，为尽量减小删除冗余词汇后对后续应用模块的不利影响，冗余词汇识别的原则是在保证准确率的情况下，尽量提高召回率。也就是说，在F值相差不大的情况下，我们选择准确率较高的方法进行冗余词汇识别。

我们把识别口语中冗余词汇的任务作为一个分类问题，通过统计分类器对口语句子中的词进行分类从而判定当前词是否是冗余词汇。

为了在冗余词汇识别过程中采用与各个冗余词汇类别相适应的特征模板信息，我们把各个类别的识别任务分开进行。例如，语气词性质的冗余词汇识别并不需要前后文的词型特征，但前后文的词型特征对于识别代词、副词、插入语性质的冗余词汇却有着积极的作用。另外，我们发现识别代词、副词、插入语性质的冗余词汇所需要的特征模板信息基本一致，为了降低识别过程的复杂性，我们把这三类的冗余成份识别任务合并在一起进行。因此冗余成份的识别任务即变为两个典型的两类分类问题：①对口语句子中的词按照语气词性质的冗余词汇prt^y和非prt^y两类进行分类。②对口语句子中的非prt^y词汇按照prt^rdp和非prt^rdp两类进行分类。冗余词汇处理的过程可用下图表示：

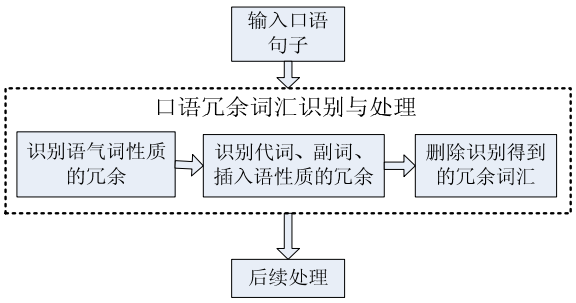

图1：口语冗余词汇识别与处理过程

[2]冗余词汇标注过程中，对于具有句型指示作用的词汇标注为非冗余词汇，以保证删除冗余词汇后，句子的句型信息不会消除。

[3] 一些长度较长的冗余现象可能并不是完全由典型的冗余词汇组成，例如“我想打听一下”，但在语料中所占比例很少[8]，且直观上对译文影响不大，因此对这部分冗余成份本文没有进行处理。

以口语句子“呃哦你这个能不能按时解决啊”为例说明这个过程。对口语句子预处理之后，首先识别句子中的语气词性质的冗余词汇“呃”、“哦”、“啊”，并对其标注为“prt^y”；然后识别句子中其他类别的冗余词汇“这个”，并标注为“prt^rdp”；最后删除标注为“prt^y”或“prt^rdp”的词，即得到句子“你能不能按时解决”，送入后续处理模块。

4. 三个统计模型简介

现有的统计分类方法中，最大熵模型近几年受到人们的重视并成功的应用于很多自然语言处理问题中，例如：词性标注、中文分词等。SVM 是一种基于统计学习理论下的数据集分割模型，它在自然语言处理领域也有着广泛应用，例如：文本分类、浅层句法分析等。而且它在两类分类问题上有着很好的分类效果。冗余词汇的识别问题还可看作一个序列标注问题，而条件随机场模型(CRFs)是序列标注模型中的佼佼者。另外，上述三种统计方法都能够融合口语句子中各种上下文特征信息而无需考虑特征之间的相关性，便于使用各种各样的特征信息，因此，本文分别使用了最大熵模型、支持向量机(SVM)和条件随机场(CRFs)三种统计方法来完成口语冗余词汇的识别任务，并选择效果最好的方法用于最终的冗余词汇识别。

4.1 最大熵模型

最大熵模型的基本思想是在满足所有已知因素的基础上建立模型,并对未知事件作最客观的估计。即在满足已知因素的条件下，选择熵最大的模型来估计数据分布[9]。使用最大熵模型进行数据分布的估计一般使用文献[10]提出的 GIS(Generalized Iterative scaling)算法，文献[11]对 GIS 算法进行了改进。本文使用最大熵模型分别完成两项任务，即：①对口语句子中的词按照语气词性质的冗余词汇 prt^y 和非 prt^y 两类进行分类。②对口语句子中的词按照 prt^rdp 和非 prt^rdp 两类进行分类。

4.2 支持向量机

支持向量机（SVM）是由文献[12]提出的一种基于两类分类问题的有监督机器学习算法，SVM 的目标是寻找一个分类面使得两类数据正确切分，且分类面距离两类数据的间隔最大化。SVM 在两类分类问题上有着良好的分类效果，因此本文使用 SVM 分别完成了两项任务，即：①对口语句子中的词按照语气词性质的冗余词汇 prt^y 和非 prt^y 两类进行分类。②对口语句子中的词按照 prt^rdp 和非 prt^rdp 两类进行分类。

4.3 条件随机场

条件随机场(CRFs)是继隐马尔可夫模型和最大熵马尔可夫模型之后提出的一个新的基于统计的序列标注模型，它在自然语言处理的领域得到了很好的应用。CRFs 是一个无向图模型，它在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布[13]。它不仅具有最大熵马尔科夫模型的优点，而且从理论上解决了标注偏置问题，达到了更好的效果。本文把冗余词汇识别问题看作一个序列标注问题，使用 CRFs 对口语句子中的词进行标注，标注类别为 prt^rdp 和非 prt^rdp。

5. 识别语气词性质的冗余词汇

语气词性质的冗余词汇是指在句子中作为冗余成份的语气词，例如“啊”、“呀”等。本部分的任务是判断口语句子中的每个词是否是语气词性质的冗余词汇。由于语气词在口语句子中出现的位置相对固定，且直观上这个问题是一个分类问题而不是一个序列标注问题，因此我们选择最大熵模型和 SVM 来完成这个任务。

5.1 最大熵模型、SVM 及特征选择

本文使用张乐博士的最大熵工具包，采用 GIS 迭代算法迭代 200 次进行训练，其他参数为工具默认值，记作 ME，特征模板如表 2 所示：

表 2：prt^y 识别—最大熵模型特征模板

特征描述	特征表示	取值
词型特征	C_0	2.5
词性特征	P_{-2}，P_{-1}，P_0，P_1，P_2	1.0
二阶特征	$P_{-2}P_{-1}$，$P_{-1}P_0$，$P_{-1}P_1$	1.0

我们使用 SVM-light V6.01 进行了两次 SVM 识别 prt^y 的实验。由于其他核函数训练时间长且效果不好，实验均采用线性核函数，其它参数为工具默认值。两次实验使用不同的特征模板：一次和上述最大熵模型一致，记作 SVM_MEfea，而另一次使用的特征模板如表 3 所示，记作 SVM_Full：

表 3：prt^y 识别—SVM_Full 的特征模板

特征描述	特征表示	取值
词型特征	C_0	2.5
词型特征	$C_i(-4<i<0$ 或 $0<i<4)$	1.0
词性特征	$P_i(-4<i<4)$	1.0
二阶特征	$C_iC_{i+1}(-4<i<3)$ $P_iP_{i+1}(-4<i<3)$	1.0
三阶特征	$C_iC_{i+1}C_{i+2}(-4<i<2)$ $P_iP_{i+1}P_{i+2}(-4<i<2)$	1.0

注：a. 表 2，表 3 中的 C 代表词型，P 代表词性，下标表示与当前词的相对位置。

b. 表 2 中取值为最大熵模型特征函数的取值，此处为经验值。实验证明这样取值实验结果更好[4]。

c. 表 3 中的取值为 SVM 特征向量的对应取值，即词型特征 C_0 对应的向量值为 2.5。

5.2 规则添加

根据对语气词性质的冗余词汇在口语句子中的使用规律的总结，在使用统计模型进行冗余词汇识别的基础上，我们又添加了以下三类规则来协助完成 prt^y 的识别，实验证明规则的添加显著提高了识别的准确率和召回率。

a) 识别口语句子中关于语气词的固定搭配，根据搭配的含义对其中的语气词进行单独处理。例如搭配“不是…吧”，对应的规则为：“吧 after 不是 notprt 5”，意思是指，如果“吧”出现在“不是”之后，并且二者之间的词的个数小于或等于 5，则判定“吧”不是冗余词汇。

b) 如果口语句子只包含语气词，用来表示应答、响应等情况，规定句子中的第一个词为非冗余词汇。该类规则是一个语气词列表，若句子中只出现这些语气词，那么就可以使用这条规则。例如：为表示“肯定答复”的含义，使用句子“啊，啊！”，则第一个“啊”标注为非冗余词汇。

c) 许多语气词的使用很简单，除了特定的结构外，在口语句子中均没有任何意义，于是都标注为冗余词汇。例如 c)类规则：“唉 prt^y”，表示只要口语句子中出现词汇“唉”，就判定为冗余词汇。

规则使用时，首先根据词的前后文信息判断是否能够使用这条规则，然后根据规则确定当前词是否是冗余词汇。例如句子“不是 没有 了 吧”。分析词汇“吧”时，由于它前面的第 3 个词为“不是”，因此可以使用规则，于是判定“吧”不是冗余词汇。

我们在上述最大熵模型和 SVM(使用与最大熵模型相同的特征模板和取值)的基础上又进行了添加规则的实验。实验过程中，首先检查当前词汇是否可以使用规则进行判定，若存在对应的规则，则根据规则对词汇进行判定。若没有，则使用统计模型对词汇进行判定。实验中共使用规则 39 条，两次实验分别记作 ME_Rule 和 SVM_MEfea_Rule。

5.3 实验

我们把本文第 2 部分介绍的语料即旅馆咨询领域 1522 句、机场信息问询领域 1471 句、以及餐馆订餐领域 1509 句合并在一起进行了语气词性质的冗余词汇的识别实验，并使用识别正确率(Precision)、召回率(Recall)和 F 值(Fscore)三项指标对实验结果进行评价。由于实验所使用的语料规模较小，我们采用交叉验证的方式进行实验以保证实

[4] 我们实验了各种不同的特征模板组合以及其它特征进行实验，实验结果都比文中的结果差。

验结果的精确性。实验过程中，我们把语料均分为 9 份，进行 9 次实验，每次实验使用其中的 8 份作为训练语料，剩余的一份作为测试语料，并将 9 次实验的结果取平均值后得到最终的结果，如表 4：

表 4：统计方法识别 prt^y 结果

模型		Precision	Recall	Fscore
最大熵模型	ME	96.02%	96.06%	96.03%
	ME_Rule	**97.45%**	96.49%	**96.96%**
SVM	SVM_MEfea	94.03%	**97.75%**	95.86%
	SVM_Full	94.04%	96.87%	95.42%
	SVM_MEfea_Rule	96.74%	97.11%	96.93%

由 SVM_MEfea 和 SVM_Full 的实验结果表明，增加的特征并没有改善 prt^y 识别的结果。从 ME 和 ME_Rule，以及 SVM_MEfea 和 SVM_MEfea_Rule 两组结果对比，我们可以看到规则的加入显著改善了 prt^y 识别的结果。另外，由 ME 和 SVM_MEfea，以及 ME_Rule 和 SVM_MEfea_Rule 的结果对比，我们发现最大熵模型结果准确率较高，而 SVM 结果的召回率相对较高，但准确率较低。由于识别冗余词汇的目的是便于后续分析，且冗余词汇识别的原则是保证准确率的情况下，尽量提高召回率，因此我们选择最大熵模型进行 prt^y 的识别。

5.4 错误分析

使用统计方法进行语气词性质的冗余词汇识别能够达到较高的正确率，但由于某些语气词的搭配结构较为复杂或者涉及到语义层面，模型很难对其作出正确的识别，主要有以下两条：

1) 语气词“啦”用于某些动词或动词短语后表示完成一种动作，不是冗余成份，但模型标注为冗余。例句子“取消啦”，“跑总站去啦”，“找到啦”等。

2)语气词“吧”、“呢”用于表示疑问语气，使得句子为疑问句型，不是冗余成份，但模型标注其为冗余成份。例如句子 “没问题吧”，“十点呢”等。

6. 识别口语句子中 prt^rdp

prt^rdp 是指在口语句子中具有代词、或者副词、或者插入语性质的冗余词汇。本部分的任务是对口语句子中的每个词按 prt^rdp 或非 prt^rdp 进行分类并进行相应标注。我们使用了最大熵模型、SVM 以及条件随机场(CRFs)进行了实验，所使用的语料为第 2 部分所得到的语料。

6.1 语料预处理

统计模型在获取上下文的词性作为特征时，不会因为词性对应的词不同而对词性特征加以区分，于是相同的词性对类别的指示作用也就相同。但冗余词汇和非冗余词汇为分类提供的信息显然不同，因此在特征采集之前，我们对词性进行修改以反应这种提供信息的差异。我们称这种策略为 PosModi.

在进行特征提取之前，我们遍历训练语料，抽取所有的 prt^rdp 冗余词汇建立 prt^rdp 冗余词词典，在之后的模型训练时，对冗余词词典中的词的词性进行修改。例如：“就”的词性为副词“d”，将其修改为：“就”词性为“d_prt”.

6.2 最大熵模型、SVM、CRFs 及特征选择

我们仍然使用张乐博士的最大熵工具包，并采用 GIS 迭代算法迭代 200 次进行训练，其他参数为默认值，记作 ME，特征模板如表 5 所示：

表 5：最大熵模型识别 prt^rdp 特征模板

特征描述	特征表示	取值
词型特征	C_0，P_0	2.5
词性特征	P_{-1}，P_1	1.0
词性特征	P_{-2}，P_2	0.5
混合特征	$P_{-1}P_0$，$P_{-1}P_1$，P_0P_1	1.0
标签特征	T_{-1}	1.0

SVM 所使用的工具为 SVM-light V6.01，由于其他核函数训练时间长且效果不好，因此我们采用线性核函数，其它参数为默认值，记作 SVM，特征模板如表 6 所示：

表 6：SVM 识别 prt^rdp 特征模板

特征描述	特征表示	取值
词型特征	C_0	2.0
词型、词性特征	$C_i(-4<i<0$ 或 $0<i<4)$ $P_i(-4<i<0$ 或 $0<i<4)$	1.0
词型、词性特征	C_{-4}，C_{-5}，C_4，C_5 P_{-4}，P_{-5}，P_4，P_5	0.5
二阶特征	$C_iC_{i+1}(-4<i<3)$ $P_iP_{i+1}(-4<i<3)$	1.0
三阶特征	$C_iC_{i+1}C_{i+2}(-4<i<2)$ $P_iP_{i+1}P_{i+2}(-4<i<2)$	1.0

我们使用工具 CRF++--0.53 进行 prt^rdp 的识别，选择的特征过滤频次下限值为 5，且由九折交叉验证取得-c 值为 0.3[5]，其他参数设为默认值，记作 CRF，特征模板选择如下：

表 7：CRF 识别 prt^rdp 特征模板

特征描述	特征表示	说明
词型、词性特征	C_0，P_0	不同模板添加两次
词型特征	C_{-2}，C_{-1}，C_1，C_2	添加一次
词性特征	P_{-2}，P_{-1}，P_1，P_2	添加一次
二阶特征	$C_{-1}C_0$，$C_{-1}C_1$，C_0C_1， $P_{-1}P_1$，$P_iP_{i+1}(-3<i<2)$	添加一次
标签特征	T_{-1}	添加一次

注： a. 表 5、6、7 中的 C，P，以及下标的含义与表 2、3 相同。T_{-1} 代表模型对前一个词的标注结果。

b. 表 5、6 中的“取值”一列的含义与表 2、3 一致。

c. 表 7 中“取值”一列的含义为对特征进行重复并添加到模型中。例如若当前词的词性为副词“d”，则构建两个模板 T00 和 T01，两个模板的取值均为“d”,并加入到模型中。实验证明，这样的方式进行特征重复能够提高识别的效果。

6.3 实验及分析

我们使用第 2 部分所得到的口语语料共 4502 句进行 prt^rdp 识别实验，并使用识别正确率(Precision)、召回率(Recall)和 F 值(Fscore)三项指标对实验结果进行评价。实验过程中，和第 5 部分采用的实验方法相同，我们把语料均分为 9 份，进行 9 次实验，并对实验结果取平均值后得到最终的实验结果，如表 8：

表 8：统计方法识别 prt^rdp 结果

模型	Precision	Recall	Fscore
ME	75.10%	78.06%	76.55%
ME+PosModi	76.60%	77.83%	77.21%
SVM	77.31%	80.84%	79.03%
SVM+PosModi	77.44%	**81.20%**	**79.28 %**
CRF	**80.48%**	77.74%	79.08%
CRF+PosModi	79.82%	77.61%	78.70%

从表中可以看出，PosModi 策略并没有改善识别结果，而且虽然各个模型结果有好有差，但总体上相差不大，结果不是很好。主要原因是所选择的特征无法真正反应 prt^rdp 冗余词汇的本质。因为无论是代词性质的冗余，还是副词性质的冗余，都是由上下文以及本身的语义关系决定，而不是它周围的词、词性等外在特征所能够单纯决定的。例如：

句(1)：你们这个是多长时间一班？

句(2)：你们这个是多长时间去一次海南？

句(1)中的“这个”指代属于句子主语“你”的某样东西，例如“大巴”。而句(2)则是说话人为维持话语连续性而加入的冗余成份。为了防止噪声信息的引入，特征选取的窗口不能太大，因此句(1)和句(2)中的“这个”的上下文信息完全一致，但却代表不同的含义。这说明无论是代词性质的冗余，还是副词性质的冗余，在相当大程度上是由长距离的上下文信息、前后句子的信息以及句子本身的语义关系决定。在满足冗余词汇识别原则的条件下，我们选择 CRFs 来进行 prt^rdp 冗余词汇的识别。

[5] -c 选项用于规定模型对训练数据的拟合程度，c 值越大，对训练数据的拟合程度越高，但拟合程度过高会出现过拟合现象。此处我们使用交叉验证的方式选择-c 的参数为 0.3。

7. 对翻译性能的影响

7.1 语料分析及评价标准

我们随机选择了旅馆咨询领域的两段口语对话 76 句、机场信息问询领域的三段对话 78 句、以及餐馆订餐领域的三段对话 76 句，共 230 句口语对话语句，使用最大熵模型识别语气词性质的冗余词汇，使用 CRFs 识别其他种类的冗余词汇并进行相应处理。

在 230 句测试语料中，共有 119 句含有冗余词汇；冗余词汇共有 205 个，占全部词汇总数的 13.9%，其中语气词性质的冗余词汇共有 141 个，占冗余词汇总数的 68.7%，其他性质的冗余词汇共 64 个，占 31.3%。

由于上述语料没有对应的翻译答案，我们使用人工主观打分的方法对结果进行评测。打分标准使用第二次国际口语翻译评测(IWSLT)所采用的打分标准，即把流畅性、充分性以及语义保持性划分为 5 个等级进行打分[14]，评价指标如表 9。

表 9：人工主观打分评价标准

等级	流畅性	充分性	语义保持性
4	完美的英语表达	全部信息充分表达出来	意思完全相同
3	较好的英语表达	绝大部分信息表达出来	意思几乎相同
2	非母语的英语表达	表达了很多信息	部分语义相同，没有误导信息
1	不流畅的英语表达	只表达了少量信息	部分语义相同，有误导信息
0	无法理解的英语表达	没有表达任何信息	意思完全相反

7.2 实验

本文分别使用目前翻译质量较好的 SYSTRAN 和 Google 翻译系统进行实验。冗余词汇处理完成后，我们使用这两个翻译系统对处理后的 230 句口语句子进行翻译。

每个翻译系统对每个汉语口语句子都会得到三句英语译文，分别对应删除冗余词汇前(记作 BeProc)，利用统计方法识别并删除冗余词汇后(记作 AfProc)，以及删除人工标注的完全正确的冗余词汇后(记作 CorrectProc)的汉语句子。

每个口语句子都由两个打分人员进行独立打分。打分完成后，对所有句子的对应指标求平均值，SYSTRAN 和 Google 两个翻译系统在删除冗余词汇前后的翻译人工打分结果如表 10，表 11：

表 10：SYSTRAN 系统翻译结果

	流畅性	充分性	语义保持性
BeProc	1.99	2.67	2.53
AfProc	2.05	2.68	2.59
CorrectProc	2.05	2.68	2.59

表 11：Google 系统翻译结果

	流畅性	充分性	语义保持性
BeProc	1.77	2.35	2.17
AfProc	1.89	2.37	2.27
CorrectProc	1.93	2.36	2.27

从表 10、表 11 中可以看出，删除冗余词汇前后，两个系统译文的充分性指标基本上没有变化，而流畅性和语义保持性指标却相应提高，这说明在对口语句子进行自动翻译时，冗余现象的存在并不会引起信息的大量丢失，但却能够影响译文的流利程度并引入误导信息。对比两个系统的 BeProc 和 CorrectProc 结果，在充分性变化不大的情况下，汉语口语句子的英语译文的流畅性和语义保持性都有了不同程度的改善，这说明预先删除冗余词汇能够改善口语的译文质量。BeProc 和 AfProc 的对比结果表明，我们的基于统计的冗余词汇识别方法能够帮助改善汉语口语的英语译文质量。

7.3 翻译结果分析

通过对删除冗余词汇前后的翻译打分结果进行分析，我们发现冗余词汇处理主要从以下两个方面改善了译文质量：

1）避免了引入误导信息

例：口语句子“奥您不订今天的是吗”

冗余处理前 SYSTRAN 和 Google 系统的翻译结果分别为：

SYSTRAN：Austria you do not subscribe today right

Google:：Austria today is that you do not set right

由于冗余词汇“奥”的存在引入了误导信息“Austria”，而删除冗余词汇后结果变为：

SYSTRAN：You do not subscribe today right

Google:：You do not set are you today

消除了误导信息。

2）使得译文较为顺畅，且能够基本保留原文语义

例：句子：“对就是樱花西街马路的西边儿”

冗余处理前 SYSTRAN 和 Google 系统的翻译结果分别为：

SYSTRAN：To is west the oriental cherry the street street west

Google:：West Street is the cherry on the west side of the street children

翻译系统把冗余词汇“就是”作为句子动词，背离了原文语义，且译文不顺畅。删除冗余词汇后的结果为：

SYSTRAN：To oriental cherry west street street west

Google:：Cherry Street on the west side of the road

冗余词汇处理后再进行翻译，译文顺畅且基本保留了源句子的含义。

删除冗余词汇之后译文总体质量得到了改善，但对于某些口语句子，冗余词汇的删除有时会导致译文质量的下降。一个原因是受翻译系统本身的影响。例如口语句子“哦就是飞机还在鄂尔多斯呢是吧”，删除冗余词汇之后的句子变为“飞机还在鄂尔多斯是吧”，两者语义完全相同，Google 系统给出的译文分别为“Oh, the aircraft is still in it, right Erdos”和“Ordos is the right aircraft is also”。虽然输入汉语句子语义相同，但冗余词汇处理前的句子翻译结果基本符合原文语义，但处理后的翻译结果却与原文语义大相径庭。这说明 Google 系统对于有些口语句子的语义无法保有忠实性，SYSTRAN 系统也有类似的问题，这里不再赘述。

另一个原因是冗余词汇有时可以作为对话子句[1][6]开始或结束的标志，也可以作为重复或修订等口语现象出现的标志。删除这些冗余词汇后会使翻译系统无法判定对话语句之间的界限，或者重复、修订等现象的出现，从而导致译文质量下降。例如：句子“到这儿可能四点吧大概”，删除冗余词汇之前 SYSTRAN 系统的翻译结果为：To here possible four probably，而删除冗余词汇之后的结果为：To here possible four general ideas。当然，Google 系统也有相同的问题出现。

对于删除冗余词汇后所导致的译文质量下降的问题，一方面需要通过发展基于真实口语对话语料的机器翻译系统来解决，而另一方面则可以通过识别标点符号位置来判定对话子句间的界限，以及将冗余、重复、修订等特殊语言现象同时进行识别处理来解决，这也将是我们下一步要进行的工作之一。

8. 结论及展望

本文对汉语口语对话语料中的冗余现象进行了分析和分类，并利用最大熵模型、支持向量机以及条件随机场模型等统计方法分别对口语句子中的冗余词汇进行了识别和处理，并取得了较好的效果。同时，本文对冗余词汇处理前后的口语句子进行了自动翻译，翻译结果证明，如果在对口语句子进行自动翻译前预先对冗余词汇进行识别处理，则能够改善口语句子的译文质量。

[6]对话子句是指一个对话语句中所包含的分句，而对话语句指从说话人开始讲话到讲完停下或被对方强行打断为止所说的全部内容。

在下一步工作中，我们将探索对于冗余词汇识别更为有效的方法以及特征信息，并考虑对口语句子中的其他特殊现象进行自动识别和处理。

参考文献

[1] 宗成庆，统计自然语言处理[M]，清华大学出版社，2008。

[2] M. Johnson, E. Charniak, and M. Lease. An improved model for recognizing disfluencies in conversational speech[C]. In Proc. Rich Text 2004 Fall Workshop(RT-04F), 2004.

[3] M. Lease and M. Johnson. Early deletion of fillers in processing conversational speech. In Proceedings of the HLT-NAACL[C], pp73-76, 2006

[4] E. Fitzgerald, F. Jelinek, and K. Hall. Reconstructing false start errors in spontaneous speech text[C]. In European Association for Computational Linguistics, pp255-263, 2009

[5] E. Fitzgerald, F. Jelinek, and K. Hall. Integrating sentence- and word-level error identification for disfluency correction[C], ACL, pp765-774, 2009

[6] Keyan Zhou, Aijun Li, Zhigang Yin, Chengqing Zong. CASIA-CASSIL: a Chinese Telephone Conversation Corpus in Real Scenarios with Multi-leveled Annotation[C]. In Proceedings of LREC, pp2407-2413, 2010.

[7] 宗成庆, 吴华, 黄泰翼等. 限定领域汉语口语对话语料分析[C].计算语言学文集(全国第五届计算语言学联合学术会议论文集), 115-122,1999

[8] 解国栋. 统计口语解析方法研究[D]. 中科院自动化所博士论文, 2004.

[9] Adam L. Berger, Stephen A.Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 22(1), pp39-71, 1996

[10] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics[J], 43, pp1470-1480, 1972

[11] S. D. Pietra, V. D. Pietra and J. Lafferty. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence[J], 19, pp380-393,1997

[12] V. Vapnik. The nature of statistical learning theory[M]. Springer. 1995

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. In Proc. of ICML, pp282-289, 2001

[14] M. Eck and C. Hori. Overview of the IWSLT 2005 Evaluation Campaign[C]. In Proceedings of IWSLT. pp11-32, 2005.

[15] 左云存, 宗成庆. 基于语义分类树的汉语口语理解方法[J]. 中文信息学报, 2006, 20(2):8-15.

[16] 解国栋, 宗成庆, 徐波. 面向中间语义表示格式的汉语口语解析方法[J]. 中文信息学报, 2003, 17(1):1-6.

作者简介：

翟飞飞(1988—)，男，硕士生，主要研究方向为机器翻译；

宗成庆(1963—)，男，研究员，主要研究方向为自然语言处理，机器翻译、口语信息处理和文本分类。