文章编号:1671-4598(2020)10-0177-04 DOI:10.16526/j.cnki.11—4762/tp.2020.10.036 中图分类号:TP391.2 文献标识码:A

智能模糊决策树算法在英语机器翻译中的应用

陶媛媛[1]，陶 丹[2]
(1. 西安交通大学 城市学院，西安 710000；2. 西安市曲江第一中学，西安 710000)

摘要：自然语言处理是计算机科学中一种从人类语言中获取和分析含义并以智能方式与人类进行交互的方法；为解决短语匹配需要目标语言中的对应语言存在偏差的问题，提出了一种基于智能模糊决策树算法的英语分级机器翻译模型（HEMTM）；模型通过搜索与分层英语机器翻译相关特征完成构建，同时，根据语言受欢迎程度和语义重要性对机器翻译的准确性进行排名，该模型在构建机器翻译的过程中，考虑了 HEMTM 与相应英语机器翻译支持关系之间的差异；研究结果显示，当采用 HEMTM 模型等级为 CFGrank 时，模型具有较高的准确性；在 $n=60$，$\delta=0$ 情况下，模型准确性为 68%；该模型可应用于具有多个答案的英语机器翻译的构建，为英语机器翻译算法领域研究提供了参考。

关键词：机器翻译；模糊决策树；分级英语机器翻译

Application of Intelligent Fuzzy Decision Tree Algorithm in English Machine Translation

Tao Yuanyuan[1], Tao Dan[2]
(1. City College, Xi'an Jiao Tong University, Xi'an 710000, China;
2. Xi'an Qujiang No. 1 High School, Xi'an 710000, China)

Abstract: Natural language processing (NLP) is a method of obtaining and analyzing meaning from human language and interacting with human in an intelligent way in computer science. In order to solve the problem that phrase matching needs the deviation of corresponding language in the target language, an English hierarchical machine translation model (HEMTM) based on intelligent fuzzy decision tree algorithm is proposed. In the process of building machine translation, the model takes into account the differences between HEMTM and the corresponding English machine translation support relationship. The results show that when the level of HEMTM model is CFG rank, the accuracy of the model is high; when n = 60, δ = 0, the accuracy of the model is 68%. The model can be applied to the construction of English machine translation with multiple answers, which provides a reference for the research of English machine translation algorithm.

Keywords: machine translation; fuzzy decision tree; hierarchical english machine translation

0 引言

自然语言处理是计算机科学中一种从人类语言中获取和分析含义，并以智能的方式与人类进行交互的方法[1]。机器翻译主要涉及使用计算机软件将文本或语音形式的语言从一种自然语言翻译为另一种自然语言，同时保留其含义和解释。从一种自然语言到另一种语言的简单单词替换是机器翻译中使用的基本过程和方法之一[2]。由于对整个短语的识别和理解，并将其与最接近的短语进行匹配需要目标语言中的对应语言存在偏差，单独使用该方法可能会导致对原始文本的误导性解释。

在机器翻译中主要部分是“翻译过程”。这个过程可以简单地解释为对源文本的含义进行解码，然后将其重新编码为目标语言。显然，此过程需要复杂的算法才能成功，因为能够完全解码示例文本的含义意味着解释器必须能够分析文本的所有功能，这就需要深入了解源语言的语法结构、语义、习语、语法等等涉及语言学的诸多方面[3]，亦不能忽略考虑源语言的文化背景。因此，正如同声传译员或者口译员需要具备大量语言学以外的知识，才能将词汇所表达的含义重新编码为目标语言，从而避免错误告知或歪曲源文本[4]。

机器翻译训练数据从来都不是完美的，双语句子对常常是错误的逐句排列，或者由于人为错误，这些句子对彼此的翻译不佳。通常，目标上下文被建模为 SMT 的语言模型。当前，主要重点工作是从单语上下文转换为双语上下文[5]，例如，双语语言模型和操作序列模型基于最小翻译单位。通常，这些方法依赖于传统 n-gram 方法，由于数据稀疏，其缺点是窗口有限且语义表示效率低下[6]。为了加强上下文的语义表示，国内外许多专家学者使用神经网络来研究相关问题（双语语境表示的神经网络）。NN 联合模型（NNJM），其编码使用前馈 NN，以减少目标方的重复发生；因此，可以集成到翻译解码中[7]。尽管如此，由于基于窗口的前馈 NN 的性质，NNJM 在捕获源侧上下文之间的长距离依赖项方面存在缺陷。

收稿日期:2020-03-10； **修回日期**:2020-03-30。
基金项目:陕西省教育厅专项科研计划项目(18JK1012)。
作者简介:陶媛媛(1986-),女,陕西商洛人,硕士,讲师,主要从事跨文化交际及英语翻译教学方向的研究。

互联网是人们获取信息的重要来源，但是互联网上存在的很多错误的分级英语机器翻译模型极大地阻碍了这一发展过程，使人们无法有效地获取信息，更无法有效的翻译信息。因此，目前对于如何建立有效的分层的英语机器翻译模型已成为迫在眉睫的问题。在互联网上，大部分的英语机器翻译的模型主要以分层英语机器翻译的形式呈现[8]。仅当信息的语义是真实的情况下，相应英语的机器翻译才是分层英语机器翻译，反之亦然，英语机器翻译在语义上是不正确的。基于分层英语机器翻译的特征，肯定存在与任何否定分层英语机器翻译平行的确定分层英语机器翻译。此外，可以通过构造相应的准确的分级英语机器翻译模型来构建否定的分层英语机器翻译模型[9]。

本文提出了一种基于智能模糊决策算法的英语机器翻译模型（HEMTM）。通过搜索有关英语机器翻译的相关HEMTM 模型来操作该模型；该模型在构建机器翻译的过程中，考虑了基于 HEMTM 与相应的英语机器翻译支持关系之间的差异。以期将该模型应用于具有多个答案的英语机器翻译的构建。

1 方法论

1.1 模糊决策树

决策树（DT，decision tree）是检索新的有趣知识的一种广泛使用方法。决策树代表了一种从标记实例中进行归纳的简单而强大的方法[10]。模糊决策树是模糊环境中决策树的推广。模糊决策树所代表的知识对于人类的思维方式来说更为自然。经典的清晰决策树广泛应用于模式识别，机器学习和数据挖掘。引入决策树来归纳分类模型，可通过沿着从根到叶的路径传播样本来对样本进行分类，该路径包含分类信息。

模糊决策树（FDT，fuzzy decision tree）是一种更通用的表示知识的方法[11]。该方法使我们能够在学习阶段（树的构造）或泛化阶段使用数字值和符号值来表示模糊模态。此外，Bouchon－Meunicr 和 Marsala 等研究人员认为模糊决策树等效于一组模糊规则并且可以引入这种归纳规则来优化数据库的查询过程或从数据中推断决策[12]。

模糊决策树的目标是具有较高的可理解性，使模糊系统具有渐进和优美的行为。因此，使用模糊集和近似推理来扩展符号决策树，以进行树的构建和推理过程。同时，借用了丰富的现有决策树方法来处理不完整的知识，并扩展为利用模糊表示中可用的新信息[12]。

模糊集的概念由研究人员 Zadeh 于 1965 年通过隶属函数提出。为了度量模糊事件，Zadeh 于 1978 年提出了可能性度量的概念。模糊熵是不确定性的一种度量。

特别地，当 ζ 是一个模糊集，取具有隶属度的值 x_i，$i=1, 2, \ldots, n$ 时，De Luca 和 Termini 分别将其熵定义为如公式（1）所示：

$$E[\zeta]=\sum_{i=1}^{n} s(\zeta=x_i) \tag{1}$$

当 $S(t)=-1\ln t-(l-t)\ln(l-t)$ 时，很容易验证该函数 $S(t)$ 关于 $t=0.5$ 对称，严格按照间隔［0，0.5］增大，严格按照间隔［0.5，1］减小，并达到其唯一最大值在 $t=0.5$ 时是 ln2。

描述熵的不确定性主要是由于语言的模糊性而不是信息的缺乏而引起的，并且当模糊变量是一个可能的变量时其消失。然而，希望看到当模糊变量退化为清晰数时熵为0，而当模糊变量为等值时熵最大。

1.2 模型构建

分层英语机器翻译的模型（HEMTM）构建如图 1 所示。输入是分层英语机器翻译，输出是分层英语机器翻译模型构建的结果。

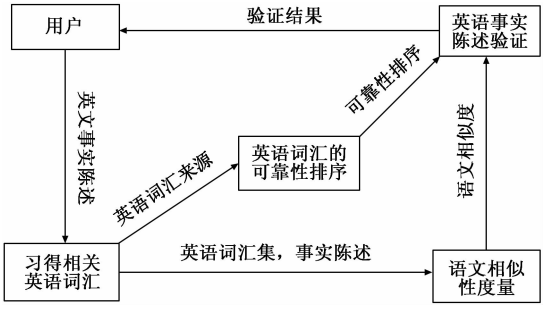

图 1 分级英语机器翻译模型

机器翻译将相关的 HEMTM 与相应的分级机器翻译相结合，为相关的 HEMTM 和相应的分级英语机器翻译之间的支持关系的评估奠定了基础。HEMTM 智能模糊决策树算法中的 r_i 和 fs 是句子的机器翻译，st_i 和 fs 是集合机器翻译[13]。词之间的机器翻译为生成语义向量和词序向量奠定了基础。单词之间的机器翻译的公式如式（2）所示。公式（2）用于计算单词的机器翻译 w_i 个和词 w_2。l 和 h 分别代表 w_1 和 w_2 在词网中的最短距离，并且 w_1 和 w_2 两者都存在于该词网。单词之间的机器翻译可以以更好的方式，通过式（2）进行评价，此时 $\alpha=0.2$ 和 $\beta=0.45$。

$$S_w(w_1,w_2)=\begin{Bmatrix} e^{-\alpha h}\cdot\dfrac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}, w_i\neq w_2 \\ 1, w_1\neq w_2 \end{Bmatrix} \tag{2}$$

在公式（1）中，如果 $w_1=w_2$，其相关性可以视为 1；此外，因为设计的词网中的信息无法覆盖所有单词。因此，如果 w_1 是个否则 w_2 无法被词网覆盖，$S_w(w_1, w_2)=0$。

假设 s_1 是句子 st_i 从 r_i 中选择的，并且 s_2 是对应的 r_i、fs 的分层英语机器翻译，接下来，将通过计算以下内容的机器翻译来演示机器翻译的过程 s_1 和 s_2。

1.3 语义向量相关性

文献［14］通过用 NN 编码整个源句子来捕获长距离依赖。此外，他们都将整个源句子在不同的翻译时间步上表示为固定向量，而不是动态向量，这在机制中已显示出了应用前景。语义向量相关性的计算：通过生成相应的句子语义向量来计算语义向量的相关性句子 s_1 和句子 s_2 以及语义向量之间的余弦机器翻译的计算。假设结束词被分为 s_1 和 s_2，相应的单词集分别是 $W_1=\{w_{11}, w_{12}, \ldots, w_{1n}$ 和

$W_2=\{w_{21}, w_{22},\ldots, w_{2n}\}$。假设 $W=W_1\cup W_2$，且 $W=\{w_1, w_2,\ldots, w_k\}$，如果 $w_i\in W_1$，那么 $vli=1$。在公式（3）中，$w_i\in W$。如果 $w_i\notin W$，并且存在最匹配的单词 w_{lm}，那么当搜索时 w_i（目标词）来自句子 s_1，然后 $vli=S_w(w_i, w_{lm})$。否则，如果 $vli=0$，将开始获取最佳匹配词的过程。

可以应用类似的计算以获得对应的语义向量 s_2，V_2。s_1 和 s_2 的语义向量相关性可以通过 V_1 和 V_2 的机器余弦转换来计算。详细的计算可以证明为式（3）所示：

$$S_s(s_1,s_2)=\frac{V_1\cdot V_2}{\|V_1\|\cdot\|V_2\|} \tag{3}$$

1.4　词序向量相关性

文献［15］引入了一种神经概率语言模型，该模型在目标语言上下文词而不是离散词的分布式表示上顺序运行。将矫正的线性单位和噪声对比估计引入 Bengio 等人的神经概率语言模，并将其应用于大型词汇。词序向量相关性的计算方法：通过生成相应的句子的词序向量，并用式（4）来计算句子的词序向量相关性，然后计算词序向量的相关性。在式（4）中，O_1 和 O_2 分别代表的词序向量 s_1 和 s_2。s_1 生成的词序向量是 $O_1=\{o_{11}, o_{12},\ldots o_{1k}\}$。结果可以通过以下方式计算：1）$w_i\in W_1$，如果 $w_i\in W_1$，o_{11} 的位置是在 s_1 中的 w_i；2）$w_i\in W_1$，如果 $w_i\notin W_1$，搜索的最匹配词 w_i，w_{lm} 已经完成。如果存在 w_{lm}，o_{1i} 的位置是位于 s_1 中的 w_{lm}，否则 $o_{1i}=0$。在找出词序向量的过程中，参数的最优值 ζ 在算法 2 中使用的是 0.4。

$$S_{re}(s_1,s_2)=1-\frac{\|o_1-o_2\|}{\|o_1+o_2\|} \tag{4}$$

1.5　智能模糊决策算法

用智能模糊决策算法计算，智能模糊决策算法 s_1 和 s_2 可以通过式（5）基于语义向量相关性和词序向量相关性来计算。如果 s_1 是句子 st_i 从中 r_i 选择，并且 s_2 是相应的英语机制翻译 fs，在式（5）中，st_i 和 fs 可以分别代表 s_1 和 s_2。在式（5）中，参数的最佳值 θ 是 0.85。

$$S(st_i,fs)=\begin{cases}\theta S_s(st_i,fs)+(1-\theta)S(st_i,fs)\\-(\theta S_s(st_i,fs)+(1-\theta)(st_i,fs))\end{cases} \tag{5}$$

上式第一个式子是 r_i 对 fs 没有倾向趋势，第二个式子是代表有倾向趋势。r_i 是否倾向于 fs 是基于获取过程中是否存在否定的语法依存关系以及否定副词在 r_i 中，例如 hardly、rarely、few、seldom 等。

2　实验分析

分层英语机器翻译模型构建的仍是当前研究热点。文献［16］使用相关语言之间的词形相似度或精确的上下文匹配来推断可能的翻译。文献［17］提出了在 ConceptNet 上的主题感知传播方法，以提高语言质量。不同的词在不同的主题下会有不同的情感。生成的主题感知情感词典提高了文本分类的性能。他们的系统预测了文本的极性以及文本中最可能的主题和概念的情感价值。文献［18］使用常识知识库来检测含义不清楚的单词。他们利用 ConceptNet 工具包确定单词替换，并计算了任意两个给定术语之间的概念相似度，并定义了平均平均概念相似度（MACS）度量标准来识别上下文外的术语。因此，本文采用的数据集是从 TREC2007 中收集的分级英语机器翻译数据集。可靠的分级英语机器翻译由 30 种，由真实语义唯一答案的分级英语机器翻译和 20 种从 TREC2007 中随机选择的多答案的真实语义的分级英语机器翻译组成[19-20]。为了进行对比分析，本实验建立了模糊算法模型（FQ）和基于模糊决策树的算法模型（HEMTM）。FQ 模型是通过搜索与分层英语机器翻译未加入特征算法的模型。实验分析了在 FQ 和 HEMTM 两种模型构建下，机器翻译的有序分布。图 2 和图 3 分别显示了当 HEMTM 数量为 150（$n=150$）时以 FQ 和 HEMTM 的模型构建方式，CBrank，CBGrank，CFrank 和 CFGrank 的分布。横坐标代表信息收集中的 HEMTM 站点，纵坐标代表相应站点中 HEMTM 的机器翻译平均排名。

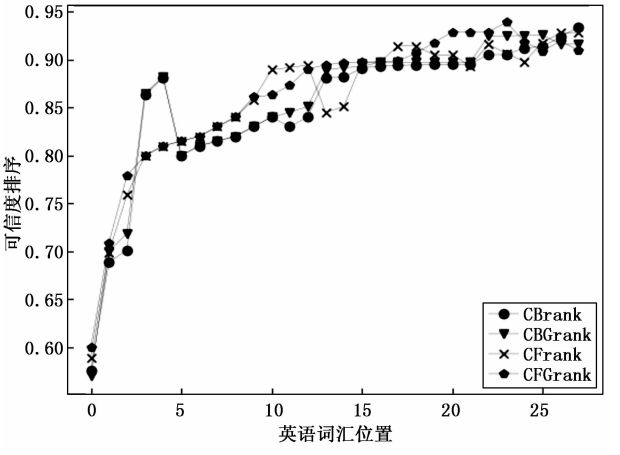

图 2　FQ 模型的机器翻译

从图 2 中可以看出，机器翻译的顺序与 HEMTM 所在的英语机器翻译信息集合的顺序没有明显的相关性。在 HEMTM 的集合中，HEMTM 机器翻译排名并不总是比质量最高的英语机器翻译排名差。究其原因，与 CBrank 和 CFrank 相比，排名间隔在 CBGrank 和 CFGrank，CBGrank 和 CFGrank 显示具有较大的跨度。可以从图 3 进行推断，机器翻译的顺序符合图 2 趋势的 HEMTM 的翻译，而在 FQ 的模型下，HEMTM 机器翻译的分布更加集中。

从上述实验中可以得出以下结论，当机器翻译等级为 CFGrank 时，构建的模型基本具有较高的准确性。图 5 描述了构建模型准确度，当机器翻译选用为 FG 模型等级为 CFGrank 时，准确度是由 n 和 δ 的关系决定。从图 4 可以看出，当 δ 是确定的时候，随着 n 的值变大，精度将上升然后下降。原因是当 n 很小时，由于相关 HEMTM 的数量有限，因此分层英语机器翻译的某些部分无法正确构建；而当 n 较大时，对相应的分层英语机器翻译的贡献率将高于对相应的分层英语机器翻译的贡献率。因此导致最后的结果为降低模型构建的准确性。而当 n 是确定的时候，精度

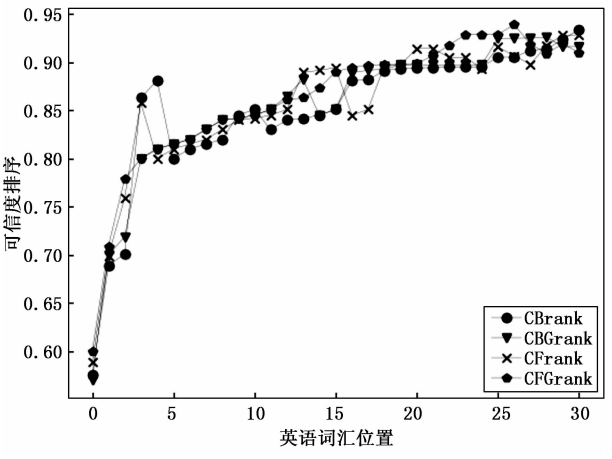

图 3 HEMTM 模型的机器翻译

将随着 δ 的增加而上升，然后再下降。

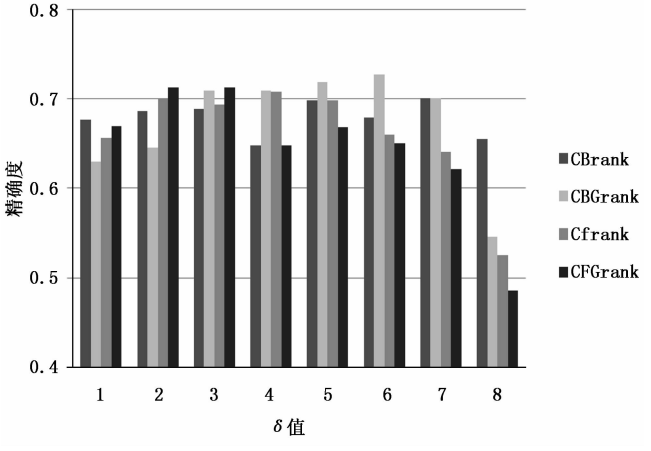

图 4 准确性趋势 $n=60$（FQ）

与机器翻译的精度不同，从该图可以看出，当 $n>90$ 时，精度随着 n 值的增加先上升然后下降。从图 4 和图 5 可以看出，当采用 FQ 的方式利用 Alexa 排名间隔的机器翻译时，可以获得较高的精度；而当对 CFGrank 进行机器翻译的排名时，可以获得更高的精度。

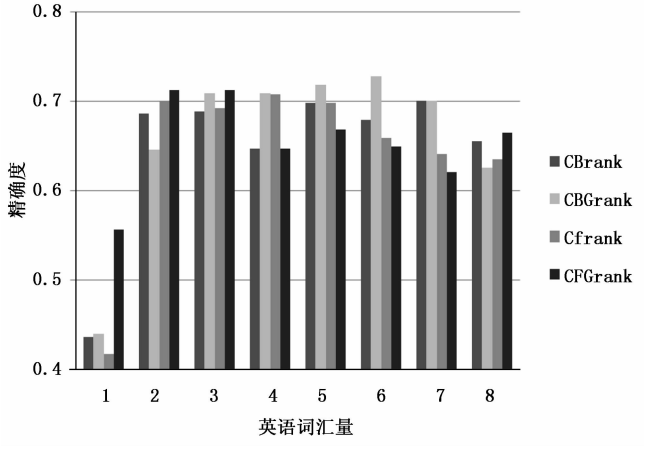

图 5 准确性趋势 $\delta=0.9$（HEMTM）

结合图 4、图 5 可以看出，4 种相关 HEMTM 模型机器翻译等级影响其准确性的参数与 FQ 模型影响准确性参数相一致。但是，由于 HEMTM 模型捕获的语言信息量质量较差，因此准确性略低于 FQ 模型。采用 HEMTM 的 CFGrank 等级，在 $n=60$，$\delta=0$ 的情况下，基本模型构建的准确性为 68%。

3 结束语

本文提出了一种基于智能模糊决策树算法 HEMTM 的分层英语机器翻译方法。通过捕获和分析相应的分层英语机器翻译中相关特征来实现模型构建。机器翻译的过程中，考虑了基于 HEMTM 与相应的英语机器翻译支持关系之间的差异。经实际验证，在 $n=60$，$\delta=0$ 时，模型准确率可达到 68%。该模型可应用于具有多个答案的英语机器翻译。

参考文献：

[1] 李艳凤. 基于直觉模糊集的决策树算法研究及应用 [D]. 北京：北京交通大学，2019.

[2] 白瑞芳. 基于 RNN 编码器的交互式机器翻译平台控制技术 [J]. 计算机测量与控制，2019，27 (7)：89－92.

[3] 倪俊杰. 机器翻译的终极之路在哪里（上）[J]. 中国信息技术教育，2020 (1)：74－78.

[4] 邵 莉. 机器翻译与本科翻译教学的困惑及展望 [J]. 海外英语，2019 (23)：54－55，59.

[5] 殷明明，史小静，俞鸿飞，等. 基于对比注意力机制的跨语言句子摘要系统 [J]. 计算机工程，2020，46 (5)：86－93.

[6] 陈祖君. 基于神经网络机器翻译模型的英文分词研究 [J]. 计算机与数字工程，2020，48 (1)：13－18，50.

[7] 黄继鹏，陈 志，芮 路，等. 基于模糊聚类决策树的分布式语者识别算法 [J]. 计算机技术与发展，2017，27 (8)：79－82，87.

[8] Chen L，Duan G. A Choquet integral based fuzzy logic approach to solve uncertain multi－criteria decision making problem [J]. Expert Systems With Applications，2020：50－56.

[9] Hao J J，Chiclana F. Attitude quantifier based possibility distribution generation method for hesitant fuzzy linguistic group decision making [J]. Information Sciences，2020：513－523.

[10] Chiara S A. Decision－making and risk in bipolar disorder：a quantitative study using fuzzy trace theory [J]. Psychology and Psychotherapy，2020，93 (1)：19－25.

[11] Aslam M，Fahmi A. New work of trapezoidal cubic linguistic uncertain fuzzy Einstein hybrid weighted averaging operator and decision making [J]. Soft Computing，2020，24 (5)：66－69.

[12] Xian S D，Yu D X. A novel outranking method for multiple criteria decision making with interval－valued Pythagorean fuzzy linguistic information [J]. Computational and Applied Mathematics，2020，39 (2)：10－16.

[13] 于振灏. 基于模糊理论的决策树算法的研究及应用 [D]. 北京：中国地质大学（北京），2017.

（下转第 185 页）