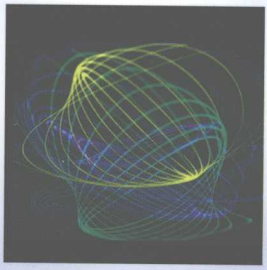
NeuroTrans: A New Model for Connectionist Machine Translation

基于人工神经网络的机器翻译

许罗迈 著



科学出版社
www.sciencep.com

NeuroTrans: A New Model for Connectionist Machine Translation

基于人工神经网络的机器翻译

[redacted] 著

NeuroTrans

A New Model for Connectionist Machine Translation

基于人工神经网络的机器翻译

许罗迈 著

科学出版社

北京

图书在版编目（CIP）数据

基于人工神经网络的机器翻译：/许罗迈著．—北京：科学出版社，2007

ISBN 978-7-03-018981-3

Ⅰ．基… Ⅱ．许… Ⅲ．人工神经元网络－机器翻译－研究－英文 Ⅳ．TP183

中国版本图书馆 CIP 数据核字（2006）第 0149849 号

责任编辑：郝建华 / 责任校对：郑金红
责任印制：钱玉芬 / 封面设计：张 放

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
中国科学院印刷厂印刷
科学出版社发行 各地新华书店经销
*
2007 年 6 月第 一 版 开本：A5（890×1240）
2007 年 6 月第一次印刷 印张：7 插页：2
印数：1—3 000 字数：265 000

定价：25.00 元

（如有印装质量问题，我社负责调换〈科印〉）

Preface

This book is based on my doctorial dissertation. It explains a new approach to machine translation with artificial neural networks. The approach treats machine translation as seeking the solution to two sub-problems. The first problem concerns obtaining vocabulary translations and is solved by using a distributed neural translation lexicon The second problem involves adjusting the transliterations produced by the lexicon into acceptable target language sentences, which is handled by a hybrid generator.

The translation lexicon learns the meaning of words from examples and stores the acquired lexical knowledge in a set of lexical networks. During translation, the lexical networks perform lexical disambiguation automatically. With the neural lexicon, programming a disambiguation component for an MT system is no longer necessary. The technique allows a translation lexicon to scale up easily to a full size lexicon for an MT application. Although developed initially for English-Chinese translation, the technique can be used to develop translation lexicons between any language pairs.

The hybrid generator consists of a generation network (GN) and a symbolic generator (SG). GN learns a simple jumble of grammar from examples, and SG physically adjusts transliterations to produce target language sentences. The generator is a bold attempt at language generation without using any formal linguistic theories. This book discusses its strengths and weaknesses. This new language generation technique is still under development, requiring further research before becoming a practical alternative to those based on the symbolic approach.

Acknowledgements

I would like to express my thanks to my supervisor Professor Gui Shichun for his understanding, tolerance and encouragement. Without this, I would not have been able to carry on my research in these uncharted waters and eventually come ashore on a newfound land, albeit not a perfect heavenly place. I also benefited from his instructions on statistics and his advice on the use of some computer packages.

I would also like to express my gratitude to Professor Miikkulainen, who took a lot of trouble to help me, a total stranger, download the source codes of DISCERN from the Internet. My thanks also go to Dr. Anguita whose MBP package and source codes were of great help in my research.

Finally, I would like to thank Dr. J. Fearon-Jones who read my draft with great care and provided many helpful suggestions. Without him, I would always feel uneasy about this book being in the hands of native English speakers.

Contents

Chapter One

Prologue

Letters of credit are important business documents in the area of international trade. These documents are mainly written in English. To ensure that a business transaction is properly handled, business people have to fully understand their content. Therefore Chinese trading companies need to translate such texts. One way to obtain such translations is by using English-Chinese machine translation packages (MT).

In the succeeding pages, we present a comparison of the translations produced by our experimental MT system NeuroTrans and two commercial MT packages. One package is Transtar v.1.0 for Windows 95 developed by China Computer Software and Technology Service Corporation (CS & S). This was first developed in the late 1980's, and the package evaluated is the latest version issued in July 1997. The other package is General Translator (通译) v.5.0 for Windows 95 and the Internet, developed by a small company and published in 1996.

We selected segments of letters of credit and translated them by using the three MT packages. The segments can be grouped into three content types. For each content type, we first present the source language text, and after it the translations of Transtar, General Translator and NeuroTrans. The comparison is shown as follows.

- **Amount expressions**

1. FOR USD 9,315 = US DLRS NINE THOUSAND THREE HUNDRED FIFTEEN ONLY
2. FOR USD 17,600.00 (SEVENTEEN THOUSAND SIX HUNDRED AND 00/100 US.DOLLARS).
3. AMOUNT: USD 2,455.00 (US DOLLARS TWO THOUSAND FOUR HUNDRED AND FIFTY-FIVE ONLY)

- **Amount translations**

Transtar

1. 对于 USD 9,315 = 美国 DLRS 九千三百十五仅仅
2. 对于 USD 17,600.00 (十七千六百和 00/100 美国。美元)。
3. 量: USD 2,455.00 (美元两千四百和 FIFTY- 五仅仅)

General Translator

1. 因为极限强度设计 9,315= 我们 DLRS 九千三百十五。唯一的。
2. 因为极限强度设计 17,600.00 (一万七千六百和 00/100。我们。美元)。
3. 合计: 极限强度设计 2,455.00 (我们美元行波振荡器千四百和。FIFTY- 五唯一的)。

NeuroTrans

1. 美元 9,315=DLRS 九千三百整
2. 美元 17,600.00 (SEVENTEEN 一万三千六百和 00/100. 美元)
3. 金额: 2,455.00 美元 (二美元二四十百 FIFTY- 五整)

- **Legal terms**

1. THIS L/C IS SUBJECT TO THE UNIFORM CUSTOMS AND PRACTICE FOR DOCUMENTARY CREDITS (1983 REVISION ICC PUBLICATION NR 400
2. THIS MESSAGE IS OPERATIVE BY ITSELF AIRMAIL CONFIRMATION

NOT TO FOLLOW CREDIT SUBJECT TO 1983 REVISION PUBLICATION 400 OF ICC

3. YOUR FAVOR IT IS SUBJECT TO THE UNIFORM CUSTOMS AND PRACTICE FOR DOCUMENTARY CREDITS (1983 REVISION INTERNATIONAL CHAMBER OF COMMERCE, PARIS, FRANCE PUBLICATION NO. 400)

- **Legal term translations**

Transtar

1. 这种 L 和碳易受于制服习惯与实践因为 DOCUMENTARY 信赖(1983 修订 ICC 出版物 NR 400)

2. 这条信息航空自己单独地是不遵循易受于 ICC 的 1983 年修订出版物 400 信任操作的证实文件

3. 它你的喜爱是易受于 DOCUMENTARY CREDITS(1983 修订国际商会,巴黎的制服习惯与实践法国出版物,号 400)

General Translator

1. 这个 L /C 工业用受统一惯例管辖的陆军和海军批准的设计实践燃料油回路附有单据的贷方(1983 修正国际计算中心发表导航雷达 400

2. 这个消息工业用技工独立航空邮寄确定不技术命令跟随,按照,从事,经营,执行贷方受 1983 修正管辖的发表 400 油浸的国际计算中心

3. 你的 FAVOR 个别训练工业用受统一惯例管辖的陆军和海军批准的设计实践燃料油回路附有单据的贷方(1983 修正国际的会议室油浸的商业,巴黎,法国发表舰艇军械 400)

NeuroTrans

1. 本信用证符合统一惯例和规范跟单信用证, 修订版 1983, 出版 400

2.本 MESSAGE 是有效 ITSELF AIRMAIL CONFIRMATION FOLLOW NOT 信用证符合 1983 修订版出版 400 国际商会

3.你方受益人受益人的 IT 符合跟单信用证统一惯例和规范(1983 修订版国际商会,

巴黎, 法国出版编号 400)

- **Payment specifications**

1.AVAILABLE WITH ANY BANK BY NEGOTIATION OF BENEFICIARY'S DRAFTS IN DUPLICATE AT SIGHT DRAWN ON APPLICANT SHOWING ISSUING BANK

2. CREDIT AVAILABLE WITH: FREELY NEGOTIABLE AT ANY BANK BY: NEGOTIATION AGAINST PRESENTATION OF THE DOCUMENTS DETAILED HEREIN AND OF YOUR DRAFTS AT SIGHT DRAWN ON FIRST INTERSTATE BANK OF CALIFORNIA, LOS ANGELES CALIFORNIA

3. CREDIT AVAILABLE BY DRAFTS AT SIGHT FOR 100 PERCENT INVOICE VALUE

- **Payment translations**

Transtar

1. 可供使用任何在中谈判的银行BENEFICIARY 的在显示的申请者上所拖拉一看见就的草案完全相同发布银行

2. 可供使用的信任:自由地可通过谈判解决在任何的银行:表达和你的草案在上所拖拉一看见就的详尽的文件的谈判在这里首先加利福尼亚,洛杉矶加利福尼亚的州际的银行

3. 一看见就百分之 100 没有声音由草案可供使用的信任评价

General Translator

1. 有效的同任何银行被议价油浸的受益人汇票磁倾角第一副本一见戴上申请人表现开证银行

2. CREDIT 有效的同:自由地可流通的验收试验任何银行被:议价反对提出油浸的文件详述在此陆军和海军批准的设计油浸的你的汇票一见戴上第一州间的银行油浸的 CALIFORNIA ,视线天使

3. CREDIT 有效的被汇票一见燃料油回路 100 百分率发票价值

NeuroTrans

1. 以下条件可议付于任何银行受益人的两份议付汇票两份即期付款人是开证人SHOWING 银行开证

2. 信用证以下条件可议付于: FREELY NEGOTIABLE 任何银行: 凭提交CALIFORNIA 银行FIRST 汇票即期单据DETAILED 在此处你方付款人是INTERSTATE, LOS ANGELES CALIFORNIA

3. 信用证以下条件可议付百分之 100 汇票即期金额发票金额

From the comparisons above, we can draw the following conclusions:

1. State of the art English-Chinese MT still has a long way to go before producing translations of acceptable quality. This is true for products of both small and larger companies. Although CS & S claims that the latest Transtar is able to produce translations with 85% readability, the performance shown above raises doubts about the stated rate.

2. Domain knowledge plays an important role in determining translation quality. It is not difficult to see that NeuroTrans produced more accurate translations for letters of credit. This is because NeuroTrans is specially designed for such a domain, although it only has a lexicon of less than 400 entries as against the larger lexicons of the other two MT systems, which have over 50,000 entries.

3. Although the comparison may seem to favor NeuroTrans unfairly, the following facts about NeuroTrans may help relieve such criticism. It has received much less human engineering effort than its counterparts. No programmer ever worked to program knowledge into NeuroTrans. Instead, it learns necessary knowledge from examples. What is more important is that the examples were

prepared not by programming professionals, but by a third-year junior middle school student. Armed with self-learning artificial neural networks, NeuroTrans could change the nature of MT development from a technique intensive job into a job for which a foreign language learner is sufficiently qualified.

4. The training of NeuroTrans took months instead of years. This indicates that MT systems catering for a large number of highly specific domains can be developed quickly. The narrowing down to specific domains is the guarantee to translation quality.

How was NeuroTrans developed? How does it differ from other MT systems? What are its strengths and weaknesses? In the rest of this book we will answer these questions. The rest of this thesis is organized as follows; in Chapter Two, we review previous MT research; in Chapter Three, we describe the theories and techniques of artificial neural networks with regard to natural language processing; in Chapter Four, we deal with the design considerations of NeuroTrans; in Chapter Five and Six, we provide the implementation details and discuss the strengths and weaknesses of the techniques we have developed; finally, in a short concluding chapter, we summarize what has been achieved in this book.

Chapter Two

MT state of the art

Machine translation is one of the earliest computer application fields people considered. After half a century of research and development, MT technology now provides only limited success within highly restricted areas. In this chapter, we will discuss the techniques that are available today and their strengths and weaknesses. In section 2.1, we present some well-known MT systems developed by using symbolic approaches. In section 2.2, we describe products which embody a compromise of the ideal and the practical: machine assisted translation. In section 2.3, we describe theories of the statistics approach to MT development, and discuss some aspects regarding the implementation of the theory. Finally, we present a summary about the merits and demerits of the two MT development approaches.

2.1 MT as symbolic systems

Most of the MT systems, whether experimental or commercial, are designed as symbolic systems. They are further divided into systems based on the transfer model or those based on the interlingua model. The transfer model seeks some direct mapping relations between the source language expressions and the target language expressions. Such mapping relations hold mainly between

the two languages involved. For translation among a set of languages, the transfer model requires $n(n-1)$ translation devices. The inter-lingua model is based on the belief that there exists a universal and language independent meaning formalism, an inter-lingua, which can be parsed from and generated into surface expressions of different languages. The model views translation as parsing a source language expression into some form of inter-lingua, and generating a target language expression from the inter-lingua. For translation among a set of languages, the inter-lingua model requires only n translation devices.

Whether using the transfer model or inter-lingua model, an MT system typically undergoes the following translating procedures (Alshawi 1994):

1. A source language string is passed to a parser which applies a grammar and a lexicon to produce a set of "deeper" syntactic/semantic forms.
2. The linguistic forms are filtered by contextual and word-sense constraints, which causes one remaining form to be passed to the translation component.
3. The linguistic form of the source language is translated into that of the target language.
4. A grammar for the target language is applied to the form and generates the target language string.

It is the linguistic forms processed at step 2 and 3 which differentiate an MT system as either a transfer or an inter-lingua system. For the transfer model, the linguistic forms take formalism closely associated with the two languages involved, and are easier for source language recognition and target language generation. For the inter-