自然言語処理 124－5
（1998. 3. 12）

# 中国の自然言語処理について

田中 康仁 †
† 兵庫大学経済情報学部
yasuhito@humans-kc.hyogo-dai.ac.jp

北 研二 ‡
‡ 徳島大学工学部
kita@is.tokushima-u.ac.jp

1997 年 8 月に、我々は中国で開催された国際会議 (標準化与技術伝播国際学術会議および語言工程) に参加し、研究発表を行ってきた。また、その際に、ハルビン工業大学を訪問し、機械翻訳、中国語入力、文字認識、音声認識等のデモンストレーションを見学した。本稿では、上記国際会議の概要および中国の自然言語処理の現状について報告する。

# Natural Language Processing in China

Yasuhito Tanaka †
† Hyogo University
yasuhito@humans-kc.hyogo-dai.ac.jp

Kenji Kita ‡
‡ Tokushima University
kita@is.tokushima-u.ac.jp

In the summer of 1997, we visited Beijing in China to attend two international conferences. The first conference was the *2nd International Conference on Terminology Standardization and Technology Transfer TSTT'97*, and the second one was the *Language Engineering*. We also visited Harbin Institute of Technology and saw demonstrations such as machine translation systems, Chinese input systems, and character and speech recognition systems. In this paper, we give a brief summary of the conferences and a survey of natural language processing research in China

## 1 はじめに

筆者 (田中) は数年毎に中国に出かけたいという思いがする。そこで、何か良い機会 (例えば学会など) がないかと考えていたところ、1997年夏2つの学会が開かれるというので、これを機会に中国の自然言語処理の研究状況を調べるためハルビン工業大学も訪れることにした。これらの計画に筆者 (北) も参加し、同行した。この内容を報告する。

## 2 国際会議と日程

**国際会議 (1) 8 月 4 日～8 月 8 日 第二届術語学、標準化与技術伝播国際学術会議**
(2nd International Conference on Terminology Standardization and Technology Transfer TSTT'97, Beijing China, 北京 中国)

この会議は中国の全国術語標準化技術審査会 (CNTCTS) と中国術語工作網 (China Termnet) と Inforterm(国際術語情報センター)ISO/TE37、ISO、IITF 等が協力して開催したものである。

簡単に説明すると、オーストリアにある Inforterm のガリンスキと中国政府の中国標準化与信息分類編纂研究所 (CSICCI) が中国の術語についての標準化の組織が固まったことと、中国が東アジアの用語術語のセンターとしての役割を果たしたというねらいから行われたものである。詳細は後で述べる。
開催場所： 中国北京, 北京国際会議場
(Beijing International Convention Hall)
**国際会議 (2) 8 月 14 日～8 月 17 日 語言工程**
(Language Engineering)

この会議は 2 年毎に開催され、中国の自然言語の研究者達と中国系で外国で研究している人々の研究発表の場である。会場は清華大学 (北京) で行われた。日本にたとえるならば言語処理学会に相当するものである。発表の言語は中国語である。

**調査 8 月 9 日〜8 月 13 日**

ハルビン工業大学訪問、自然言語の研究状況について調べた。筆者の強い要望で北京大学の俞士汶教授の紹介でハルビン工業大学に行くことが実現した。ハルビンはロシアにも近いし、戦前は日本の満州国のあった所でもあり、機械翻訳に関心の深いところである。

## 3 国際会議の内容について

**TSTT'97(2nd International Conference on Terminology Standardization and Technology Transfer**

この国際会議は用語の標準化と用語の移転についてのものである。

この会議の最初の 4 つの講演は、以下の通りである。

1. Terminology Market, Christian Galinski.
2. Making Great Effort to Create Chinese Terminology with National Feathure of China Giving More Contribution to Terminology in the World, X. U. Jiakc.
3. Terminology Standardization in China Review and Project, JIAO Yunqin.
4. Environmental Terminology Development on WWW – A Progress Report, Gerhard Buclin.

そして 5 つのセクションがあった。これらは順番に 2 日間で行われた。

**Section II** Computer-Aided Terminology
13 論文、発表論文 11 編

**Section III** Termbank and Its Applications
13 論文、発表論文 10 編

**Section IV** Lexicography and Terminology
6 論文、発表論文 4 編

**Section V** Terminological Activities for Minority Language in China
10 論文、発表論文 4 編

集まった論文は 100 編を越えるものがあったが、その中から 78 編を論文集に入れ、40 編に発表を許した。しかし、数人は欠席したため、30 数編と最初の 4 つの講演であった。

日本からの発表は以下の 5 件であった。

- Self organization of multilingual terms for innovative use of information
  藤原 譲 (神奈川大学)
- Problems in the standardization of terminology for computer education anf for multiple cultures
  張志良 (上海師範大学)、伊吹公夫 (東京工科大学)、余錦華 (東京工科大学)
- Creation of a terminological database from a multilingual indexing vocabulary
  山下泰弘 (学術情報センター)、春山暁美 (専門用語研究会)、久保田 均 (労働省産業医学総合研究所)
- A mutual reference retrieval system for Japan/China-MARC using NDC and LC
  川手太士、石川徹也 (図書館情報大学)
- Language activity and information processing
  田中康仁 (兵庫大学)

太田は欠席者の時間のところで発表した。

- The harmonization of food terms in East Asia and a proposal for preparing a multilingual food thesaurus throughout East Asia
  太田康弘 (文教大学)

なお、日本からの参加者は 11 名であった。

この会議での印象であるが、言語や用語を取り扱うことを専門とする社会が確立しはじめたといってよい。

中国語内の問題としては大きな問題が 2 つあり、その 1 つは台湾、香港での用語と中国大陸の用語に相異が発生しはじめている点である。科学技術と経済的活動が活発な情報、通信、インターネットの分野と物理学 (半導体関係、電子部品等) の用語である。

もう 1 つの問題は中国は多くの少数民族を持っている。モンゴル族、チベット族、朝鮮族、… 等があ

る。これらの民族言語をうまく処理できるシステムを作らなければならない。固有の文化、文字も持っている。もし、これを単純に漢民族の方針に統一しようとすれば、大きな反乱や政治問題にも発展する。

この会議を主催しているオーストリアのガリンスキ氏の考えは、言語の政策や理論に主体をおき、応用やコンピュータを利用しての実際的なものはあまり好まないという考えであるが、今回の発表をみると政策、理論、実用が三分の一ずつ配置されている感じであった。実用について少し配慮する面が出てきたようである。これは筆者(田中)と日本人参加者の感想である。

中国は、政府の機構として中国標準化信息分類編纂研究所を設立し、用語の標準化と用語の移転についての本格的研究を行い始めたことは興味深いことである。日本に於いても、このような研究所がほしいものである。

### 言語工程 (Language Engineering)

全国等四届計算語学総合学術会議は、8月14日～8月17日まで北京清華大学で行われた。113編の論文が集まり62編が選ばれた。この会議では、次の7つの分野での研究発表が行われた。

(1) 中国語の句や語について 11 編

(2) コーパスやコーパスの作成、加工技術について 9 編

(3) 基本語のコーパスや言語分析方法

(4) 文の分析と生成 7 編

(5) 機械翻訳 8 編

(6) 文の検索、自動的文の検索、文の校正 14 編

(7) 人工知能型文字入力方法と人と機械のヒューマンインターフェイス 8 編

その他12編は概要が論文集に入っている。田中と北は次のような内容で発表を行った。

- 帯有対訳信息的慣用表示的収集
  田中康仁, 穂志方, 俞士汶
- 基於概率模型的語言的集類方法: 根据多語種語料庫進行語言系統樹的再構造
  北研二, 穂志方, 俞士汶

これからも分かるように筆者等は中国語ができないので北京大学の大学院の博士課程の穂志方さん(女性)に訳と発表を依頼した。質問にはそれぞれ英語で答えた。

筆者(田中)は中国語で自己紹介し、中国の黒龍江省ハルビン市で生まれ、牡丹江で育ったことと、戦後、撫順に6ヶ月滞在し、大連経由で日本に帰ったことを英語と中国語(地名のみ)で行った。中国の人は大変驚いていた。

中国の自然言語処理は、ここ6～7年で急速な進歩をとげている。この理由には次のようなことがある。

1. パーソナルコンピュータ、ワークステーションが大学の研究室では自由に使えるようになったこと。
2. 中国語の基礎的辞書が北京大学の俞士汶教授らによって作られ安く配布されている。
3. 人民日報等の機械可読データやコーパスが出まわっている。
4. 開放政策のおかげで経済が発展し外国との貿易、交流がさかんになり言語処理に対する需要が出ている。例えば中英、英中、日中、中日の機械翻訳が出ているし、ソフトウェア会社がかなりの精度のものを作り販売を開始していること等がある。

日本の経済成長が外国の技術導入により急速に進展したのと同じ側面が中国の発展にある。しかし、今後、意味処理、文理解等について同じように進むかどうかはわからない。しかし、12億の中から選ばれた研究者であるからもっと進むかもしれない。

## 4 ハルビン工業大学

ハルビン工業大学では計算機学科の李生教授と王教授の2つの研究室を訪問した。

李生教授は副校長(日本での副学長)でもある方で機械翻訳については中国でも有名な1人である。

この研究室では次のような研究が行われており、これらのデモンストレーションを見学した。

1. ルールベースの機械翻訳システム (中↔英)
2. 用例ベースの機械翻訳システム

3. 中国語と英語のコーパスの中から用語を指定し、中国文と英語を検索するシステム

4. 中国文の朗読システム

機械翻訳システムは開発が終り、これから実用化にむけてのレベルアップの段階であった。サンプルを入力すると良い翻訳結果であった。しかし、問題点も多々あるようであった。

中国文の朗読システムはできあがっていたが、少し不自然な点があるようであった。今後は中国南部の広東語にも適用したいようである。

王教授研究室では3つの研究が行われていた。

1. 中国語入力システム

2. 中国語の文字認識システム

3. 音声認識システム

中国語の入力システムは連続してピンインを入れると、個々の漢字のピンインの曖昧さが減少するという性質を利用し、単語、句の入力を簡単にするシステムである。

中国語の文字認識システムは携帯型の装置上に文字を書くと認識し、幾つかの文字候補の中から最適なものを選択するものである。

音声認識システムは高雑音下で約200語程度の語彙を認識するものであった。

これらの研究成果の一部は米国の大手ソフトウェア会社が実用化しているものもある。また、日本の企業が入力システムとして売り出しているものもある。

このほかハルビン工業大学の他の学科のロボットの研究室や精密加工の研究室等も見学した。また大学全体の見学も行った。博士課程500人、修士課程1500人、学生15000人、教職員全部合わせると2万5000人程度の学校であった。

李生教授は副校長でもあるので、我々を公式の訪問者として本館の応接室で外事部の教授(女性)等と会談した。徳島大学が姉妹校となっていたことは大変良いことであった。

## 5 機械翻訳のソフトウェア会社について

TSTT'97の国際会議、語言工程会議で入手したパンフレット等により中国の機械翻訳のソフトウェア会社について述べる。

### 中国計算機軟件与技術服務総公司仲軟訳星公司

この会社はCICCプロジェクトの成果を基にして機械翻訳のシステムを作り、日本の会社が「中日自動翻訳ソフト」TRANSTARとして売り出している。価格49,800円であり、7万2千語の基本語辞書を用いA4用紙1ページ/30秒で翻訳可能である。
日本の会社は、

システム日本サイエンス(株)
〒221 横浜市神奈川区反町2-16-2 マックサワトビル2F
Tel. 045-324-2115 Fax. 045-324-7394

### 北京高立公司

この会社は中国の機械翻訳の草わけであった黒龍江大学機械翻訳研究室の成果を引きつぎ日中、英中の開発を行っている。7万語の基本辞書と10分野の専門用語辞書70万用語を持っている。このほかあと12分野の専門用語も拡大する予定である。翻訳率は英中が80%程度、日中は70〜75%程度である。英中デモンストレーションを見学することができた。中国語が良いかどうかは判断がむずかしいが見学している人達の反応は良かった。スピードも問題なかった。英中は既に1,000ユーザに売っているとのことであった。6〜7年前から行っており、6〜7名が開発を行っている。

### 天津大通通訳計算機軟件研究所

この会社は天津市にあり、1992年頃から機械翻訳の研究および開発を行っている。

英中、中英の機械翻訳システムの開発を行っている。約40人で開発している。ユーザは1万件ほどである。翻訳精度は85%程度である。翻訳速度は3万字/時である。専門用語は20分野で総計200万語持っている。今後4つの分野も拡大しようとしている。インターネット上の翻訳システム、OCR使用の翻訳システム、個人用、業務用等もある。将来的には韓国語、ロシア語、日本語、独語にも拡大したいようである。

英語から中国語へのサンプルを図1に示す。

television is a system for transmitting moving pictures over long distances by means of radio .
电视是系统为了发送活动图画在长的距离上用无线电.
in some respects , television and radio are alike .
在一些方面.电视和无线电是相似的.

図 1: 英語から中国語への翻訳例

このほか、日本の企業で中国語機械翻訳システムの開発を行って売り出している企業がある。

**クリエイト大阪**

この会社は大連理工大学の傅教授と一緒に日中、中日の機械翻訳システムを作成し「そんごくう」という名で売り出している。中国語の入力システムも販売している。

1997年初夏、中国の大連市にも関係会社を作り開発を行っている。

**高電社**

英日、日英の機械翻訳システムをパーソナルコンピュータ上で実現し、安く売り出したことで有名である。ここでは韓国語の機械翻訳システムも販売しているし、中国の自然言語研究者と協力し、中国語のシステムも開発している。

**その他**

日本の大手企業も最近では中国語の入力システム、翻訳にも興味を示し始め、中国の大学と提携し研究開発を進めている。

## 6 北京市とハルビン市

北京は数年毎に大きく変わっているようである。古い住宅の建物はなくなり、新しいビルがどんどん建設されている。高速道路もでき空港と北京市は近くなった感じがする。市民の足も自転車から自動車、タクシーへと移り変わっている。

北京市の周辺部に多くのビルが建設されているのが印象的であった。開放政策のせいか市民生活にもゆとりが感じさせられるようであった。

ハルビン市は第二次大戦前は白系ロシア人(ソビエト革命の時に追われた資本家が亡命した人)や日本人が作った都市といっていい。そのためか市の中心部にはロシア建築を思わせるものがあった。しかし、戦後50年もたつと建物も古くなるため改築したり、取りこわして新しく作っている。しかし、昔の建物に復元しているものもある。

ハルビンの有名な百貨店秋林は戦前三越によって作られたもので、今でも高級品が売られていた。また、隣の大きなビルは立派な本屋で、専門用語辞書を十数冊筆者(田中)は買った。

また、ロシア正教会の大きな円形ドームの建物は改修され、そのまわりは公園として整備されようとしていた。ハルビンはすっかり生まれかわろうとしている。

ハルビンはきれいな都市で筆者(田中)の生まれた所でもあるので公式日程のあいまに見学した。

## 7 おわりに

中国の自然言語処理の研究は急速ないきおいで進み、辞書、コーパス、各種自然言語処理の道具が充実しはじめている。

また、自然言語に対する需要からソフトウェアハウスが、中国語入力、エディタ、機械翻訳システム、検索システム等を開発し販売を開始していることがわかった。数年後にはもっと進み、色々な自然言語の道具が一般企業、生活の場へと普及しはじめていることだろう。これは筆者達の感想である。

## 参考文献

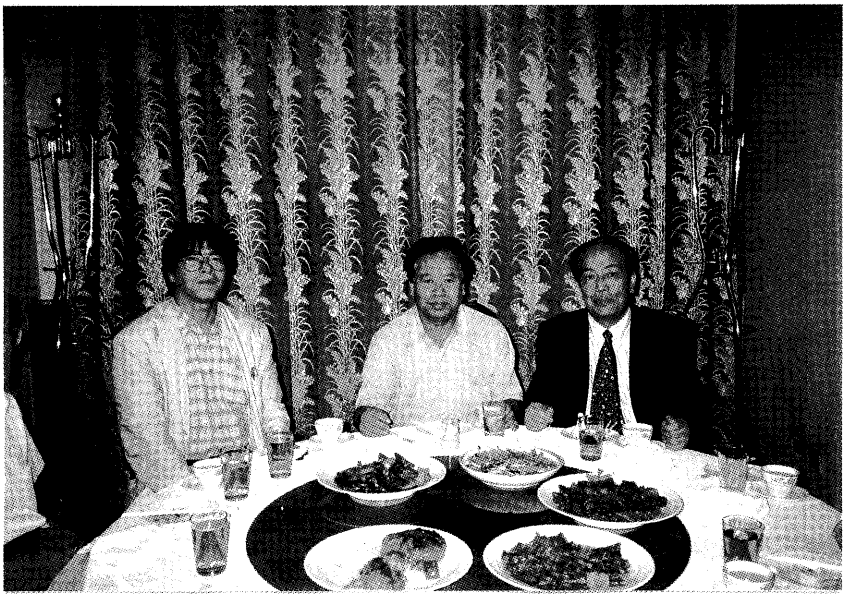
[1] 語言工程, 精華大学出版社.

[2] 2nd International Conference on Terminology Standardization and Technology Transfer, TSTT'97 Proceedings, 中国大百科全書出版社.

# 付録：中国での研究者達との写真


標準化与技術伝播国際学術会議にて

北京大学 俞志汶教授 (左) と一緒に



ハルビン工業大学 李生教授 (中央) と一緒に

白系ロシア美人と一緒に