2007－NL－177 (1)
2007／1／26

決定木を用いた中国語の疑問文の訳語選択ルールの生成

古宮嘉那子 高虹 但馬康宏 小谷善行
東京農工大学 工学部 電子情報工学専攻

Abstract

中国語の疑問文における「怎么」は，英語に翻訳すると，how や why など，複数の結果が得られる．本論文では，「怎么」の使用法によってどのような訳語が適切であるかを決定する中英翻訳システムを作成し，その選択ルールを用例により決定木学習を用いて自動的に作成した．

決定木の学習および評価に用いるデータは人手で 309 件の文と，属性，および結果を作成したものである．本研究での決定木は二分木であり，決定木学習のアルゴリズムには C4. 5 を用いた

実験の結果，最高値で 99. 7%の正解率で語を選択することが可能となった．また，作成した木により，自動的に訳語選択ルールを生成した． 作成した木は，ヒューリスティックな決定木をほぼ再現した．

Generating a Set of Rules to Determine the Translation of a Chinese Word "怎么 (zenme)" Using Decision Tree Learning

Kanako Komiya, Gao Hong, Yasuhiro Tajima and Yoshiyuki Kotani

Department of Computer, Information and Communication Sciences

Tokyo University of Agriculture and Technology

Abstract

A Chinese word "怎么 (zenme)" in questions has some translations such as "how" and "why". The authors made a Chinese-English translation system to determine the suitable translation of "怎么 (zenme)" in questions according to the usages. It generates a set of rules to determine them automatically, by decision tree learning.

Training data and testing data are manually made. They are 309 sentences and their features, values and the results. The authors used binary tree and C4.5.

The accuracy of the system was 99.7% and the decision tree derived from experiments is almost the same as the decision tree that are made manually.

1． はじめに

同じ表現を使用しても，その使用法によって，その求められる訳語の表現が異なることがある．[1]は英日翻訳において，多義語「Take」を，決定木学習を用いて適切な日本語にする研究を行っている．また，[2,3]では状況により適した敬語を選択するルールを生成しており，このルールは状況に適した敬語の訳語を選択するルールとしても利用できる．

中国語の疑問文における「怎么」は，英語に翻

訳すると，how や why など，複数の結果が得られ，それらの使い分けが行われている[4]．本論文では，「怎么」の使用法によってどのような訳語が適切であるかを決定する中英翻訳システムを作成し，その選択ルールを用例により決定木学習を用いて自動的に作成した．

2． 中国語の疑問文における「怎么」

中国語の疑問文における「怎么」は，英語に翻訳すると，how や why など，複数の結果が得られる．

[4]によれば，怎么は疑問代名詞であり，疑問文において述語(predicate)または副詞修飾語（adverbial modifier）として使われる．怎么？怎么了?怎么啦?という形で大抵表されるものが，述語としての使用例であり，What's wrong? What's the matter? What's up? と訳される．副詞修飾語としての使用例はもっと多く，why, how, What do you mean by saying? や，怎么一回事？または怎么回事？という形でWhat's the matter? と訳されるものがある．これらの役割を引用して列記すると，why の役割である理由，how の役割である方法，修辞疑問，What do you mean by saying?の役割である説明要求，What's wrong? What's the matter? What's up? の役割である，何があったか尋ねるなどがある．これらの表現は，主に怎么の後置語によるものである．

[4]では，文型と役割および訳語の関係として，以下 a～i のような結論が得られている．左の部分は文の構造を表し，それに対して右の部分でその役割と，訳語が指定されている．

a. 怎么+副詞+動詞 <理由> why
b. 怎么+代名詞+動詞/形容詞
<理由> why
c. 怎么+没/没有 <理由> why
d. 怎么+動詞+不+補語 <理由> why
e. 怎么+動詞
<理由> why
<方法> how
<説明> What do you mean by saying?
f. 怎么+能/可能/可以…(呢)
<修辞疑問> how
g. 怎么+个+動詞+法 <方法> how
h. 怎么+个+副詞+法
<説明> What do you mean by saying?
i. 怎么+个+形容詞+法
<説明> What do you mean by saying?

3． 訳語選択システム

訳語選択システムでは，まず，ステージ 1 の決定木作成フェーズにおいて，ユーザは訳語選択システムに学習データを入力する．この学習データは，訳語を選択するための要因と，その状況における訳語によってなる．訳語選択システムは決定木学習を行い，訳語選択ルールを作成する．

次に，ステージ 2 の実行フェーズにおいて，ユーザは訳語を選択するための要因，つまり状況を入力する．訳語選択システムはステージ 1 の決定木作成フェーズで得られた訳語選択ルールを用いてその状況に適した訳語を選択する．（図 1）

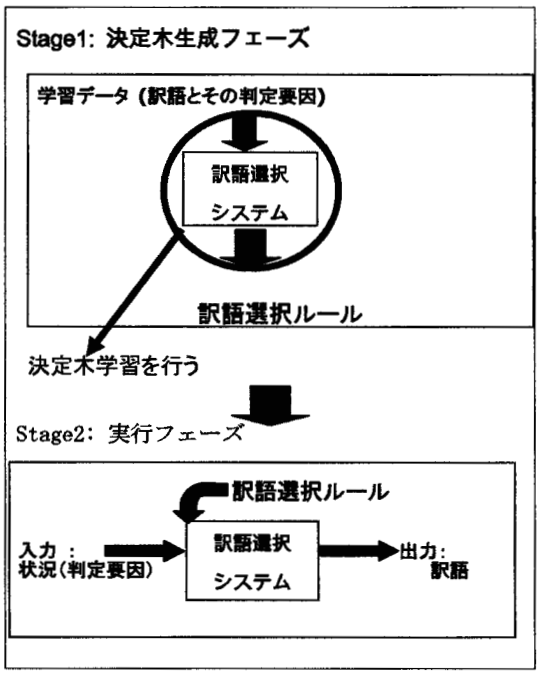

図 1 訳語選択システム

訳語選択システムは用例により自動的にルールを作成して，それを基に適切な訳語を選択してくれるというものであり，どのような原因でどのような語を訳語として使うかということが明示的に分析できる利点を持つ．また，自動的に得たルールにより訳語の用法について，細かい使い分けを自動的に行ってくれる．

4． 訳語とその判定要因

本章では，訳語選択システムの訳語とその判定要因について述べる．本章の1節では，訳語選択システムのデータについて述べ，2節では，訳語選択システムの判定要因について述べる．最後に，3節では，訳語選択システムにおける訳語について述べる．

4.1 訳語選択システムのデータ

ここではシステムを作りやすくするために，この研究で扱う文を定義する．中国語の疑問文における「怎么」は，英語に翻訳すると，howやwhyなど，複数の結果が得られる．本論文で扱う訳語選択システムは，中国語の疑問文における「怎么」を，適切なルールを生成し，選択するものである．ひとつの出力に対する入力はひとつの文についての複数の情報であるが，扱うすべての文において，「怎么」はひとつであるとする．さらに，文のうち，関係ないと思われる部分は省略可能であるとする．なお，学習データは辞書（[5,6]）を参照したものと，人手で作成したものからなっている．

対象となる文は形態素解析されているものとする．また，ここで適切な訳語を決定する要素は既に分かっているものとする．使用法の分類は人手で行う．

4.2 訳語選択システムの判定要因

本節では，訳語選択システムで用いる訳語の判定要因の選択肢を明らかにし，それらを選択する基準について述べる．本節で述べる訳語選択システムの判定要因は，訳語選択システムの決定木作成の際に与えられる学習データの一部となり，訳語選択システムの入力となるものである．

まず，本論文では，[4]のルールを採用し，これに以下の四つのルールを加えた．ここでも，左の部分は文の構造を表し，それに対して右の部分でその役割と，訳語が指定されている．

j. 怎么?/怎么了/啦?/怎么[一]回事?
<何があったか尋ねる> What's wrong?
What's up? What's the problem?
k. 怎么+動詞+得+補語
<修辞疑問> how
l. 怎么+前置詞句+動詞
<方法> how （<理由> what）
m. 怎么+動詞+得+代名詞(这么/这样/那么/那样)+形容詞
<理由> why

なお，文献[4]によれば，aの文は起こるべきではないのに起こってしまったことについて尋ねるのに用いる用法であり，副詞+動詞の理由について尋ねており，動詞の理由について尋ねているのではない．

また，fとkの文は修辞疑問を表しているが，kは特にその状況が起こりえないことを表している．jとkは両者ともhowと訳すのが適当であり，全体の文としては，how can...として訳すのが適当である．さらに，kの呢は形式的な語なので，省略可能である．

そして，lの文は文脈により，howとwhyが決定される．道理に叶ったことならばhowが，あり得ないことに対してはwhyが選択される．用例の数はhowの方が多い．

さらに，動詞を分類することにより，細かい規則を発見した．その分類とは，Va，Vb，VcとVtである．ここで，Vaは，“+了”で理由を尋ね，“+的”で方法を尋ねる動詞であり，方向を持った動詞である．

また，Vb は理由を尋ねる動詞であり，その内訳としては like や regret にあたるような精神的な行為を表す動詞，cheat や pity にあたるようなある一定の方向を指す行動を表す動詞，seem や be にあたるような連結を行う動詞，lack や increase にあたるような増減を示す動詞，love や advise などその他の動詞がある．Vb のうち，Vt は，ある一定の方向を指す行動を表す動詞である．Vc は方法を尋ねる動詞であり，以上のもの意外を指す．

以上の a～l の規則により完成したヒューリスティックな決定木は以下となる．

1.怎么?怎么[一]回事?怎么了/啦?
　→what’s wrong?
　what’s the matter?
　what’s up?
　what’s the problem?
2. 怎么の後置語が副詞　→why
3. 怎么の後置語が代名詞(这么/这样/那么/那样)
　→why
4. 怎么の後置語が没/没有　→why
5. 怎么の後置語が能/可能/可以　→how
（可能の後に語がなければ，how come?となる）
6. 怎么の後置語が个
　その次の語が動詞
　→how
　その次の語が副詞（好/不好/容易/难…）
　→ What do you mean by saying?
　その次の語が形容詞
→ What do you mean by saying?
7. 怎么の後置語が動詞
　動詞の種類が Va
　了で終わる文の場合　→ why
　的で終わる文の場合　→ how
　動詞の種類が Vb
　Vt+你
　→ What do you mean by saying?
　それ以外　→ why
　動詞の種類が Vc
　V+这种+n → why
　それ以外　→ how
　その次の語が不+補語　→ why
　その次の語が得
　その次の語が補語　→ how
　その次の語が代名詞+形容詞　→ how
8. 怎么の後置語が前置詞句+動詞
　→ why　または　how

以上から，訳語選択システムにおける状況は，以下，1)～7)のように設定した．

1) 怎么の直後の品詞
1. adverb
2. preposition
3. 没 / 没有
4. 能 / 可能 / 可以
5. 个
6. verb
7. prepositional phrase
8. その他

2)動詞の種類
問題となる疑問文で，どのような動詞が使われているかによって，訳語が別れる．そのため，動詞を 5 種類に分類した．
1. Va: “+了”で理由を尋ね，“+的”で方法を尋ねる動詞
2. Vb: 理由を尋ねる動詞
3. Vc: 方法を尋ねる動詞
4. Vt: Vb の一部であり，ある一定の方向を指す行動を表す動詞
5. W: What do you mean by saying などに対応

3) 怎么の後の後の品詞
1. verb
2. adverb
3. adjective

4. 你
5. 这种
6. 不
7. 得
8. その他

4) 最後に含まれるものは何か
1. 了
2. 的
3. その他

5) 怎么？怎么啦/了?怎么[一]回事? かどうか（例外処理）
1. yes
2. no

6) 可能の後に語はあるかどうか
1. no
2. yes

7) 得の直後の品詞
1. preposition
2. complement
3. その他

4.3 訳語選択システムにおける訳語

本節では，訳語について述べる．これらは，訳語選択システムにおいて，決定木作成の際に与えられる学習データの一部であり，訳語選択システムの出力となる．

ここで訳語選択システムで扱う訳語は，以下の5種である．

1. What's wrong? / What's the matter? / What's up? etc
2. How come?
3. why
4. how
5. What do you mean by saying?

5. 訳語選択システムにおける決定木学習

決定木は，属性と結果により構成される木であり，一つのノードは属性によって結果を分類する．決定木学習とは，この木の性質を用いた機械学習であり，学習データを与えて木を生成し，生成した決定木をルートノードからトップダウンで辿ることで，適切な結果を選択することができるようにする手法である．本システムでは，ノードに使用法を格納しておき，葉の部分に適切な翻訳結果（訳語）を格納した決定木を用いる．なお，決定木の生成には，貪欲法に含まれる，C4.5[7]を用い，決定木は二分木を生成した．

6. 訳語選択システムの実験

五分割交差検定によって精度を見る．閾値は以下の二通りを利用する．

- エントロピー
- エントロピー×データ件数

付録の表1に，全データ309件の訳語の種類とその件数およびパーセンテージを示す．

なお，この実装した決定木を309件の学習データに対して，クローズドデータで実験したところ，100%の正解率を得た．また，このときの葉の数は13であった．

7. 訳語選択システムの結果およびその評価

訳語選択システムの，実験の閾値と正解率と葉の数を調べた．葉の数により結果の数が異なるため，葉の数により正解率が変化する．そのため，できるだけ葉の数に偏りが出ないように実験を行った．

付録の表2と表3に，本論文における訳語選択システムの，実験の閾値と敬語表現の優先順位別の，正解率と葉の数を示す．表において，それぞれの列の最高値は太字で表示している．さらに，全体の最高値は，下線で示した．

実験の結果，閾値が0のとき，最高値99.7%の正解率となり，308件が正解した．最高値が閾値を0にしたときであるため，訳語選択システムで

は，閾値による最良の正解率の差はないと言える．

このときに生成した決定木5つのうち，もっとも正解率の良かったものを図2として巻末の付録に添付する．

なお，この決定木データのうち，最高の割合を占める「why」の割合が56.63 %であるため，実質的には43.05ポイントの正解率向上が認められる．

8. 訳語選択ルールの考察

本章では，実際の例を通して，訳語選択ルールについて考察する．

例1)「怎么？」

この例の訳語の判定要因を，以下に示す．

怎么の直後の品詞 :その他　**動詞の種類** :動詞なし **怎么の後の後の品詞**:それ以外
最後に含まれるものは何か :それ以外　**怎么？怎么?/了?怎么[一]回事? かどうか** :yes
可能のあとに語があるかどうか :yes　**得の直後の品詞** :それ以外

これは，以下のように決定木をたどって，「What's wrong? etc 」を返した．正解である．「怎么の直後の品詞 = その他」の場合，必ず「What's wrong? etc 」になっているため，常に正解になっている．

IF (怎么の直後の品詞 = その他)　yes <-
What's wrong? etc　4件

例2)「怎么又买衣服了?」

この例の訳語の判定要因を，以下に示す．

怎么の直後の品詞 : 副詞　**動詞の種類** : Va **怎么の後の後の品詞**: 動詞
最後に含まれるものは何か : 了　**怎么？怎么?/了?怎么[一]回事? かどうか** : no
可能のあとに語があるかどうか :yes　**得の直後の品詞** :それ以外

これは，以下のように決定木をたどって，「why」を返した．正解である．Yes をえらべる要因がないため，残りのデータとなっていることが推察できる．

IF (怎么の直後の品詞 = その他)　no ->
IF (可能のあとに語があるかどうか = no)　no ->
IF (怎么の直後の品詞 = 个)　no ->
IF (怎么の後の後の品詞 = あなた)　no ->
IF (得の直後の品詞 = 補語)　no ->
IF (怎么の直後の品詞 = 能／可能／可以)　no ->
IF (最後に含まれるものは何か = 的)　no ->
IF (怎么の後の後の品詞 = それ以外)　no ->
why　95件

図2に示した決定木を見てみると，ヒューリスティックな決定木をほぼ再現していることが分かった．また，最高の深さは11であり，シンプルな決定木になっている．データを人出て作ったため，作為的な結果になってしまったが，その分精度はとてもよかった．

9.まとめ

「怎么」の使用法によってどのような訳語が適切であるかを決定する中英翻訳システムを作成し，その選択ルールを用例により決定木学習を用いて自動的に作成した．

実験の結果，最高値で99.7%の正解率で語を選択することが可能となった．また，作成した木により，自動的に訳語選択ルールを生成した．作成した木は，ヒューリスティックな決定木をほぼ再現した．

参考文献

[1]水野秀紀，荒木健治，栃内香次： 決定木アルゴリズムを用いた多義語の訳語選択手法の有効性の評価, 情報処理学会研究報告, 98-SLP-24-3, pp.17-24 (1998) .

[2] 古宮嘉那子，小林明子，乾伸雄，小谷善行： 決定木学習による敬語の選択ルールの生成　――敬語選択システム――，情報処理学会　第67回全国大会（平成17年）　公演論文集　第二分

冊 1ZA-3, pp429-430 (2005).
[3] Kanako Komiya, Yasuhiro Tajima, Nobuo Inui, Yoshiyuki Kotani, Generating a Set of Rules to Determine Honorific Expression Using Decision Tree Learning, Lecture Notes in Computer Science, Volume 3878, Jan 2006, pp 315 – 318 (2006).
[4] Wu YunFang, Lexical Disambiguation of "怎么"in Questions in Chinese—English Machine Translation.《辉煌二十年—中国中文信息学会二十周年学术会议》北京：清华大学出版社, pp 263-271 (2001).
[5] 北京・商務印書館 小学館 共同編集：日中辞典, 小学館, 1998 初版 17刷.
[6] 愛知大学中日大辞典編纂処編：中日大辞典, 日中大辞典刊行会, 1978 5刷.
[7] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, (1993).

付録

表 1 訳語の種類とその件数およびパーセンテージ

訳語の種類	件数[件]	パーセンテージ [%]
What's wrong? , What's the matter? What's up? etc	5	1.62
How come?	2	0.65
why	175	56.63
how	90	29.13
What do you mean by saying	37	11.97

表 2 訳語選択システムのエントロピーを閾値に用いた際の正解率と葉の数

エントロピー	0	0.5	0.7	0.8	0.9	1	1.2	1.5	2
葉の数	12.8	10.4	8	7.2	5.4	5	4.2	1.6	1
正解率[%]	**99.68**	94.19	89.33	83.84	71.85	70.56	67.96	56.97	56.64

表 3 訳語選択システムのエントロピーにデータ件数をかけた値を閾値に用いた際の正解率と葉の数

エントロピー ×データ件数	0	50	100	150	200	250	300	350	400
葉の数	12.8	9.8	7.8	6	5	4.2	4	2.6	1
正解率[%]	**99.68**	92.24	87.38	78.32	70.56	67.96	66.03	58.58	56.64

0 可能のあとに語があるかどうか=No

<--- yes 葉 1 how come? 2 件

no ----> 2 怎么の直後の品詞=その他

<---- yes 葉 5 What's wrong? etc 4 件

no ----> 6 怎么の直後の品詞=个

<---- yes 13 怎么の後の後の品詞 =動詞

<---- yes 葉 27 how 10 件

no ----> 葉 28 What do you mean by saying 22 件

no ----> 14 怎么の後の後の品詞 =你

<---- yes 29 動詞の種類=Vc 以上未満 (Vt か否か)

<---- yes 葉 59 What do you mean by saying 8 件

no ----> 葉 60 why 1 件

no ----> 30 得の直後の品詞=それ以外

<---- yes 葉 61 how 19 件

no ----> 62 怎么の直後の品詞=能／可能／可以

<---- yes 葉 125 how 17 件

no ----> 126 最後に含まれるものは何か=的

<---- yes 葉 253 how 9 件

no ----> 254 怎么の後の後の品詞 =それ以外

<---- yes 509 動詞の種類=Vc

<---- yes 1019 怎么の直後の品詞=没／没有

<---- yes 葉 2039 why 3 件

no ----> 葉 2040 how 17 件

no ----> 葉 1020 why 39 件

no ----> 葉 510 why 96 件

図 2 作成した決定木