

Improving Web Search with Language Technologies

Thomas Hofmann

Director of Engineering - Zurich





Improving Web Search with Language Technologies

- 1 Lexical Semantics
- 2 Machine Translation
- 3 Information Extraction
- 4 Automatic Speech Recognition



1 Lexical Semantics

Improving Ads Targeting & Search Quality



Natural Language Processing for Search Quality

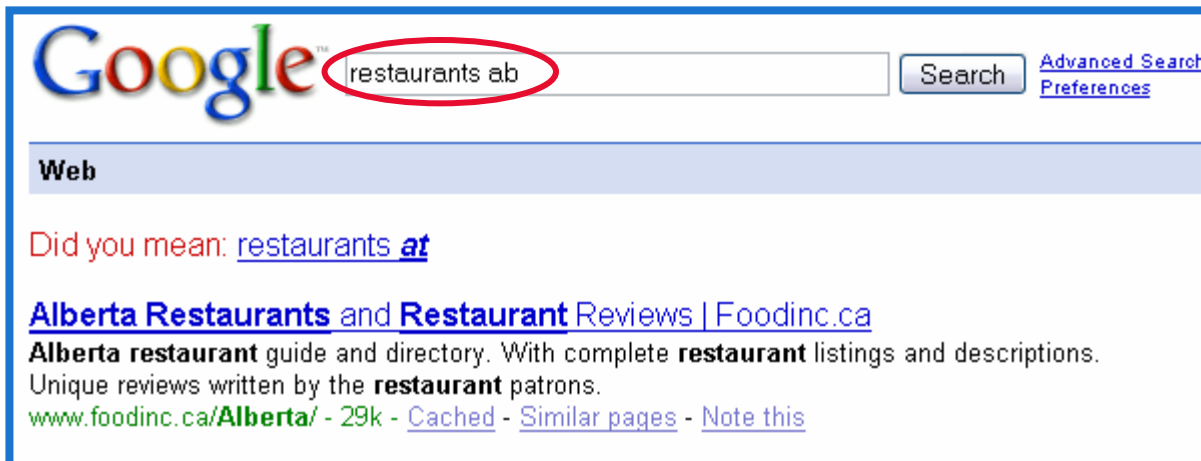
Two main ingredients: **stemming** and **synonyms**

Challenges for **synonym** expansion

- **Learning** of lexical semantics from data
- High **precision** in order to avoid loss of topicality
- Use **context cues** to trigger synonyms

Natural Language Processing for Search quality

Synonym expansion depends on context:



A screenshot of a Google search interface. The search bar contains the text "restaurants ab", with "ab" circled in red. To the right of the search bar is a "Search" button and links for "Advanced Search" and "Preferences". Below the search bar, the "Web" section is highlighted. The first result is "Alberta Restaurants and Restaurant Reviews | Foodinc.ca", with a description: "Alberta restaurant guide and directory. With complete restaurant listings and descriptions. Unique reviews written by the restaurant patrons." and a URL: "www.foodinc.ca/Alberta/ - 29k - Cached - Similar pages - Note this".

ab = Alberta



A screenshot of a Google search interface. The search bar contains the text "ab 1492", with "ab" circled in red. To the right of the search bar is a "Search" button and links for "Advanced Search" and "Preferences". Below the search bar, the "Web" section is highlighted. The first result is "Industrial Control - Screw Connection Terminal Blocks", with a description: "Allen-Bradley's Bulletin 1492-J line of internationally approved IEC style terminal blocks offers a wide range of features and benefits ideally suited for ..." and a URL: "www.ab.com/en/epub/catalogs/12768/229240/229268/3170951/2297059/ - 25k - Cached - Similar pages - Note this".

ab = Allen
Bradley



Expanded Matching in On-line Ads Targeting

Targeting mechanisms for **AdWords**: match user queries with advertiser (bidded) keywords

Types of matches

- **Phrase match**: all tokens from a keyword appear consecutively in the query, and in the same order
(keyword) used cars -> (query) cheap used cars
- **Broad match**: all tokens from a keyword appear somewhere in the query, regardless of order
(keyword) used cars -> (query) used toyota cars
- **Expanded broad match**: some tokens from a keyword or its related words appear in the query
(keyword) used cars -> (query) used automobiles, automobiles



Expanded Matching in On-line Ads Targeting

[Advanced Search](#)
[Preferences](#)

Web

Results 1 - 10 of about 1,290,000 for **used automobiles**. (0.18 seconds)

[Used Cars from Japan](#)

www.tradecarview.com

Great news for importers + dealers! Wide range of vehicles from Japan.

Sponsored Link

Sponsored Links

[Used Cars, New Cars, Buy a Car, Sell a Car, Car Dealers ...](#)

Search over 3 million new and **used** cars, find new and **used** car dealers, sell a car, and research your car purchase. More people buy cars at AutoTrader.com.

www.autotrader.com/ - 37k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Export **used** cars from USA](#)

Participate in **Used** Car Auctions Membership only \$150. Join today

www.ExportTrader.com

[Used Automobiles including **used** cars at CarsDirect.com](#)

Used automobiles, **used** trucks and **used** convertibles are for sale at CarsDirect.com.

www.carsdirect.com/help_center/buying_advice/used_automobiles - 35k -

[Cached](#) - [Similar pages](#) - [Note this](#)

[Automobile](#)

Trova gratis la tua auto tra oltre 1.500.000 annunci

www.AutoScout24.it

[New Car Prices, **Used** Car Pricing, Car Reviews by Edmunds Car ...](#)

Edmunds car buying guide lists new car prices, **used** car prices, car comparisons, car buying advice, car ratings, car values, auto leasing.

www.edmunds.com/ - 67k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Auto](#)

Tutto l'Usato da tutta Italia.

Trova subito la tua nuova auto!

www.secondamano.it/Auto_Moto

[New Car Prices | **Used** Car Values - Official Kelley Blue Book Site](#)

New car prices and **used** car values from Kelley Blue Book. New Car Blue Book values. ...

Used cars for sale. Sell my car. Car insurance. Auto loans.

www.kbb.com/ - 90k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Used trucks & trailers](#)

Stock: 1200 **used** vehicles. Holland.

All brands, all types, check our www

www.KleynTrucks.com

[Auto](#)





2 Machine Translation

Enriching Web Content



Machine Translation for Web Search

Machine translation system developed in-house at Google
(Franz Och)

Goals: enrich Web content in languages with limited content

Usage: Web page translation, translate this page link on result page, cross-language retrieval (Russian, Arabic)

Challenges in machine translation:

- MT from English into other **target languages**
- MT for any **text types & topics**
- Model size **optimization** & efficient **search**
- Interface, usability, user feedback





Text and Web

[Translated Search](#)

[Dictionary](#)

[Tools](#)

Translate Text

Original text:

Spanish to English



Translate

Translate a Web Page

http://

Spanish to English



Translate

[Google Home](#) - [About Google Translate](#)

©2008 Google





[Text and Web](#)

Translated Search

[Dictionary](#)

[Tools](#)

[Help](#)

Translated Search

Search for: Translated to: نور دبي - [Not quite right? Edit](#)

My language: Search pages written in:

Translated results from Arabic web pages

Results 1 - 10 of about 4,760 for نور دبي.

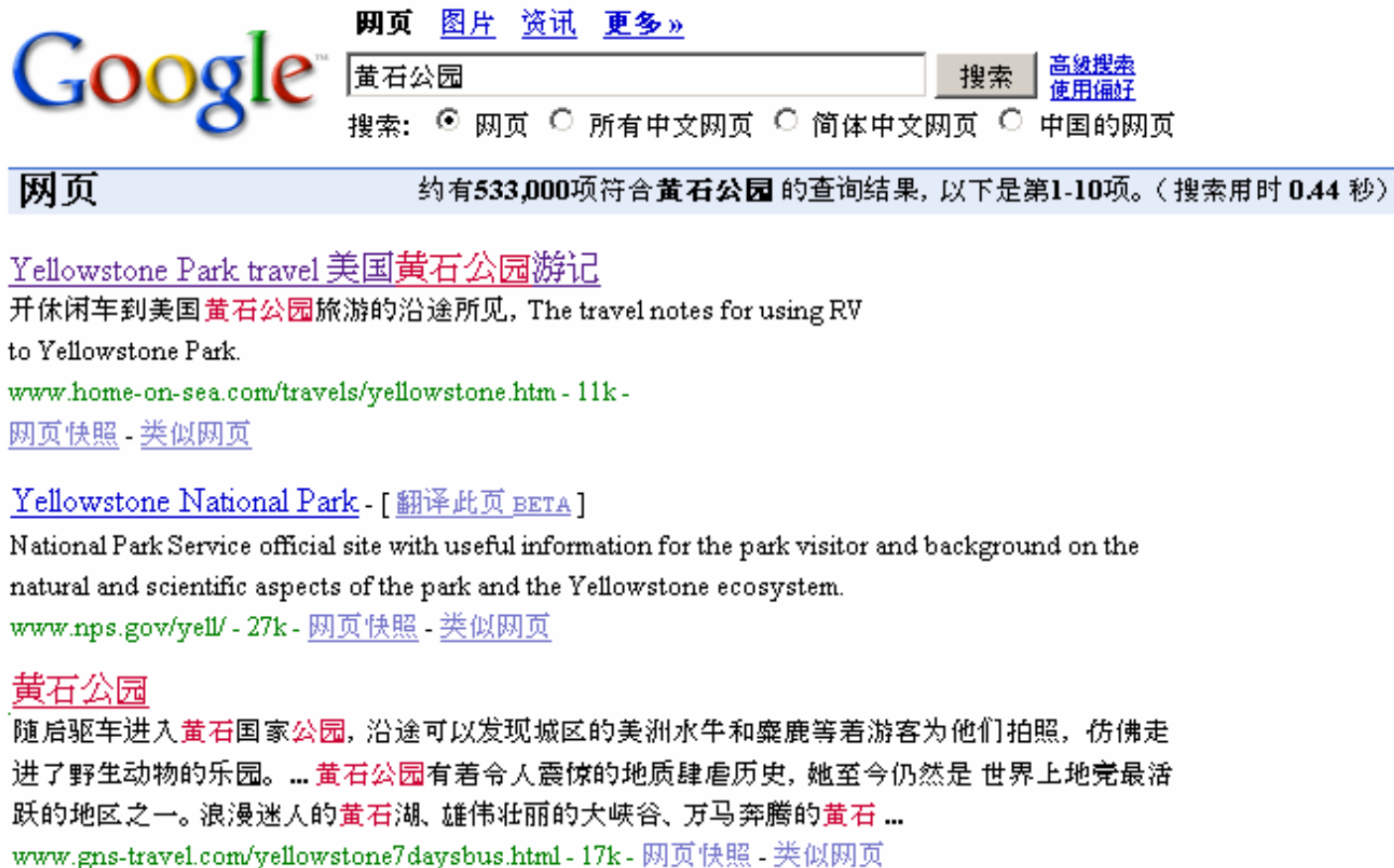
English translation

[Believe me I Amowooowoooot!!!! -- Nem - girls - a world Eve...](#)
Mckowooworh darlin **TOURS Dubai** ... Old 06-30-2006, 01:38 PM...The reason bad guys unless renewed vigor gives you the evil sister **TOURS Dubai**.
Signature. Vrack break my back...
www.nua3m.net/vb/forum4/thread2565.html - 151k - [Cached](#)

Original Arabic - [Hide Arabic results](#)

[صندفوني إنتي أموووووووووت!!!! - نواعم - بنات - عالم حواء ...](#)
مشكووووووره حبيبي نور دبي ... قديم 06-30-2006, 01:38 PM ... والسبب رفاتي السوء اللهم اكفنا شرهم
بخطبك الحافيه اخي نور دبي. التوقيع. فراقك كسر ظهري ...
www.nua3m.net/vb/forum4/thread2565.html - 151k - [نسخة مخبئة](#)

Search Results – “Translate this page” link



Google [网页](#) [图片](#) [资讯](#) [更多»](#)

黄石公园 [高级搜索](#)
[使用偏好](#)

搜索: 网页 所有中文网页 简体中文网页 中国的网页

网页 约有**533,000**项符合**黄石公园**的查询结果, 以下是第**1-10**项。(搜索用时 **0.44** 秒)

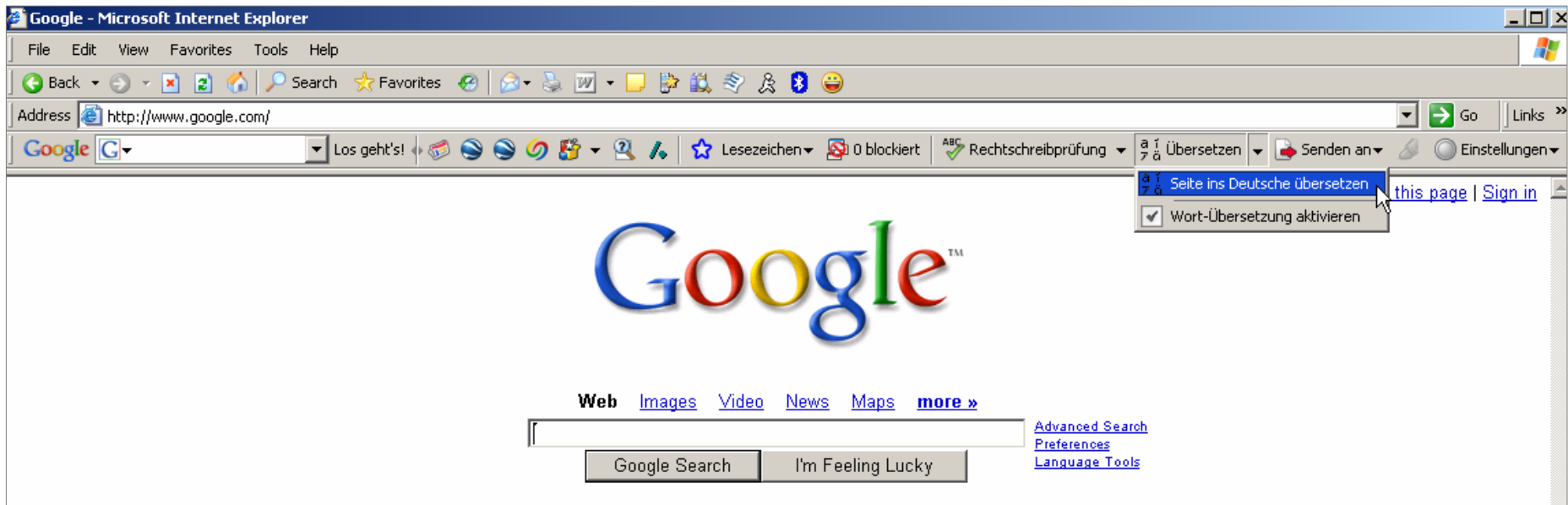
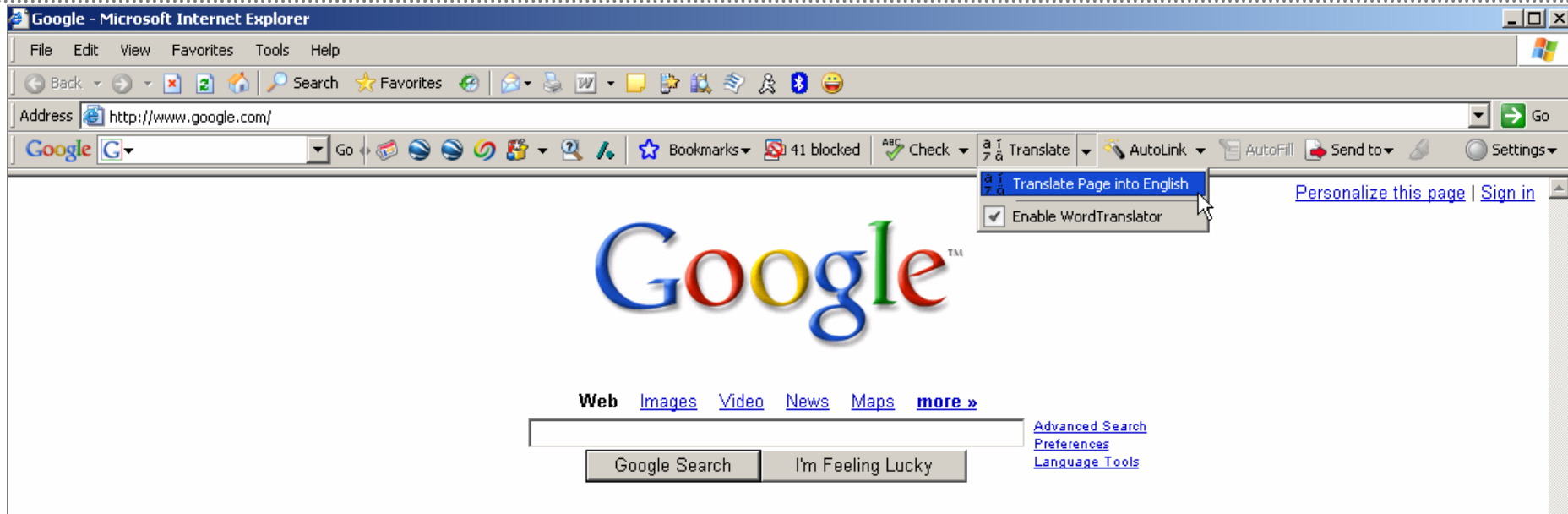
[Yellowstone Park travel 美国黄石公园游记](#)
开休闲车到美国**黄石公园**旅游的沿途所见, The travel notes for using RV to Yellowstone Park.
www.home-on-sea.com/travels/yellowstone.htm - 11k - [网页快照](#) - [类似网页](#)

[Yellowstone National Park](#) - [[翻译此页 BETA](#)]
National Park Service official site with useful information for the park visitor and background on the natural and scientific aspects of the park and the Yellowstone ecosystem.
www.nps.gov/yell/ - 27k - [网页快照](#) - [类似网页](#)

[黄石公园](#)
随后驱车进入**黄石国家公园**, 沿途可以发现城区的美洲水牛和麋鹿等着游客为他们拍照, 仿佛走进了野生动物的乐园。... **黄石公园**有着令人震惊的地质肆虐历史, 她至今仍然是 世界上地壳最活跃的地区之一。浪漫迷人的**黄石湖**、雄伟壮丽的大峡谷、万马奔腾的**黄石** ...
www.gns-travel.com/yellowstone7daysbus.html - 17k - [网页快照](#) - [类似网页](#)



Translation in Google Toolbar



Translation Feedback -- Launched in Feb '07

STANFORD UNIVERSITY

网站 人 go

史丹福指数址

二 三 四 五 六 八 我

十 钾 1 米 叙 澳 叙 住

收 笔 吴 五 x 肤 的 宅

News

2006年11月14日

X射线激光脉冲捕捉图像:科学家们对首次使用极短激烈X射线激光脉冲获得高分辨率纳米尺度物体形象之前雷射摧毁样本。

2006年11月14日

沟通气候变化:树林环境研究所本月推出跨大学学者集训,以增进

Original English text:

The Cherry Orchard : Anton Chekhov's final play, with a new translation by Marina Brodskaya.

[+ Suggest translation](#)

Events

<櫻桃園:安东契诃夫最后扮演着新匠brodskaya翻译, 11月16日至18日上午8时许,在戏剧工室罗布莱。

程

福

心

动

网站 人 go

史丹福指数址

二 三 四 五 六 八 我

十 钾 1 米 叙 澳 叙 住

收 笔 吴 五 x 肤 的 宅

News

2006年11月14日

X射线激光脉冲捕捉图像:科学家们对首次使用极短激烈X射线激光脉冲获得高分辨率纳米尺度物体形象之前雷射摧毁样本。

Original English text:

The Cherry Orchard : Anton Chekhov's final play, with a new translation by Marina Brodskaya.

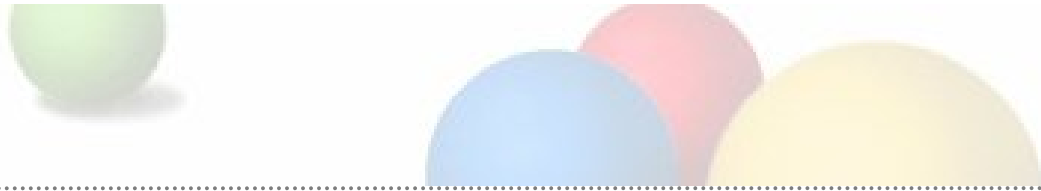
[- Suggest translation](#)

<?桃园 安?契?夫最后扮演着新匠brodskaya翻?.

Clear Submit

Events

<櫻桃園:安东契诃夫最后扮演着新匠brodskaya翻译, 11月16日至18日上午8时许,在戏剧工室罗布莱。



3 Information Extraction

Supporting Question Answer Retrieval

Information Extraction for Question-Answer Retrieval

Open domain **extraction of facts** from the Web

Goals: provide succinct answers to queries that are questions

Usage: currently triggers a special “search onebox” to deliver a fact

Challenges in information extraction:

- **Reliability** of extracted facts
- **Coverage** of relevant facts from all domains
- **Reputation** of sources and combination thereof
- Triggering of Q&A retrieval
- Combination of evidence and inference



Question Answering Retrieval: Example

Compile fact with source reference for simple question-like queries:



The image shows a screenshot of a Google search interface. At the top, the Google logo is on the left, followed by a search box containing the text "Population of Japan". To the right of the search box is a "Search" button and two links: "Advanced Search" and "Preferences". Below the search bar, a blue header bar indicates "Web" on the left and "Results 1 - 10" on the right. The first search result is highlighted with a red border and contains the following text: "[Japan](#) — **Population:** 127,433,494 (July 2007 est.) According to <https://www.cia.gov/library/publications/the-world-factbook/print/ja.html>". Below this, there are three more search results, each with a blue underlined title and a snippet of text followed by a green URL and blue links for "Cached", "Similar pages", and "Note this".

Japan — **Population:** 127,433,494 (July 2007 est.)
According to <https://www.cia.gov/library/publications/the-world-factbook/print/ja.html>

[Population](#)
The **population** under 15 years of age in **Japan** was 17.4 million as of April 2007 ... The speed of aging of **Japan's population** is much faster than in advanced ...
www.stat.go.jp/English/data/handbook/c02cont.htm - 20k - [Cached](#) - [Similar pages](#) - [Note this](#)

[POPULATION OF JAPAN](#)
2000 **Population of Japan** provides statistical data on the current states of **population**, number of and structure of families and the structure of industries ...
www.stat.go.jp/english/data/kokusei/2000/final/hyodai.htm - 69k - [Cached](#) - [Similar pages](#) - [Note this](#)


[Japan Geography](#)
The **population of Japan** is about 125000000, including approximately two million foreign residents. More than half of the non Japanese **population** is of ...
www.japan-guide.com/list/e1000.html - 29k - [Cached](#) - [Similar pages](#) - [Note this](#)





4 Automatic Speech Recognition

1-800-GOOG-411



Automatic Speech Recognition

1-800-GOOG-411 service from mobile phones

Goals: local business information completely free, directly from your phone

Usage: easy to use speech interface for mobile devices

Challenges:

- **Speaker variability**
- **Background noise**
- **Navigation & usability**

I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.
 I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.
 I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.
 I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.
 I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.
 I will use Google before asking dumb questions. I will use Google before
 asking dumb questions. I will use Google before asking dumb questions.

