Why Wikipedia Needs to Make Friends with WordNet

Kow Kuroda*, Francis Bond*,** and Kentaro Torisawa*

*Laguage Infrastructure Group, MASTAR Project, NICT, Japan

**Nanyang Technological University, Singapore

Enthusiasm for Wikipedia

- Wikipedia is a dream of a resource with very broad coverage.
  - There are a number of enthusiasts of Wikipedia in NLP.
  - It is regarded as a triumph of *Collective Intelligence* (Levy 1997; Tovey (ed). 2008)
- Some of them claim that WordNet (Fellbaum, ed. 1998) and the like are dispensable if we have Wikipedia.
  - They typically criticize (i) narrow coverage of terms and (ii) subjectivity of sense identification.

But wait

- How grounded is such a claim?
  - Is broader coverage always preferable over higher precision?
    - Precision of automatic term recognition affects the result we get.
    - It can be good for segmented languages but it is not true for unsegmented languages like Japanese. Errors in the stage of tokenization/morphological analysis lowers precision drastically.
  - Is everything written in text, in the first place?

NiCT National Institute of Information and Communications Technology

MASTAR

Question and Answer

- Question
  - Is WordNet dispensable if we have Wikipedia?
- Our tentative answer is *No*.
  - More precisely, it is not true unless high-precision automatic term recognition and term abstraction is achieved.

National Institute of Information and Communications Technology

MASTAR

Outline of talk

- Report issues experienced in the construction of hypernym hierarchies from 2.4 million hypernym-hyponym pairs (Sumida et al. 2008).
  - pairings over 95,000 hypernym tokens and 0.9 million hyponym tokens (including notational variants)
- Report results from comparison of elements in the hypernym hierarchies thus constructed against lemmas of Japanese WordNet (Bond et al. 2008, 2009).
- Conclusions

NiCT National Institute of Information and Communications Technology

MASTAR

Construct hypernym hierarchies from Japanese Wikipedia by Gradual Term Abstraction (GTA)

NiCT National Institute of Information and Communications Technology

MASTAR

Relation acquisition from the Wikipedia

- Sumida et al (2008) proposed a method of automatically acquiring hypernym-hyponym relations from the Japanese Wikipedia.
  - They used Support Vector Machines (SVM) (Vapnik 1995), one of the most powerful machine learning techniques.
- With the 90% precision threshold, 2.4 million hypernym-hyponym pairs were acquired.
  - 2.4 million is an impressive number well beyond personal productivity.

NiCT National Institute of Information and Communications Technology

MASTAR

Problems

- Acquired pairs are not clean enough and not as useful as expected because
  - Automatic relation extraction suffers a lot from errors at the term extraction/recognition stage.
    - This is more serious in unsegmented languages.
  - Even if extraction is successful, the result needs to be mapped onto existing ontologies effectively.
- This requires Gradual Term Abstraction (GTA).

NiCT National Institute of Information and Communications Technology

MASTAR

Gradual Term Abstraction

Why is it necessary?

- Given the observation that a large number of hyponyms acquired from the Wikipedia denote named entities, GTA of their hypernyms should produce mapping from them to upper ontologies.
- GTA is useful because such lower-level hypernyms are referred to as instances of compound noun phrases, and they can be linked to lexical databases like WordNet as they stand.

Gradual Term Abstraction
What is it?

- Suppose we have a hypernym-hyponynym pair (famous British rock singer, Peter Gabriel).
- GTA is a task where
  - a specified term (e.g., *famous British rock singer*) is gradually converted into less specified ones (⇒ *British rock singer* ⇒ *rock singer* ⇒ *singer*) by removing modifiers one by one.
- In theory, GTA of term set *T* in language *L* automatically produces links it to upper ontologies for *T* if WordNet of *L* is provided.

NiCT National Institute of Information and Communications Technology

MASTAR

Gradual Term Abstraction
How it is performed

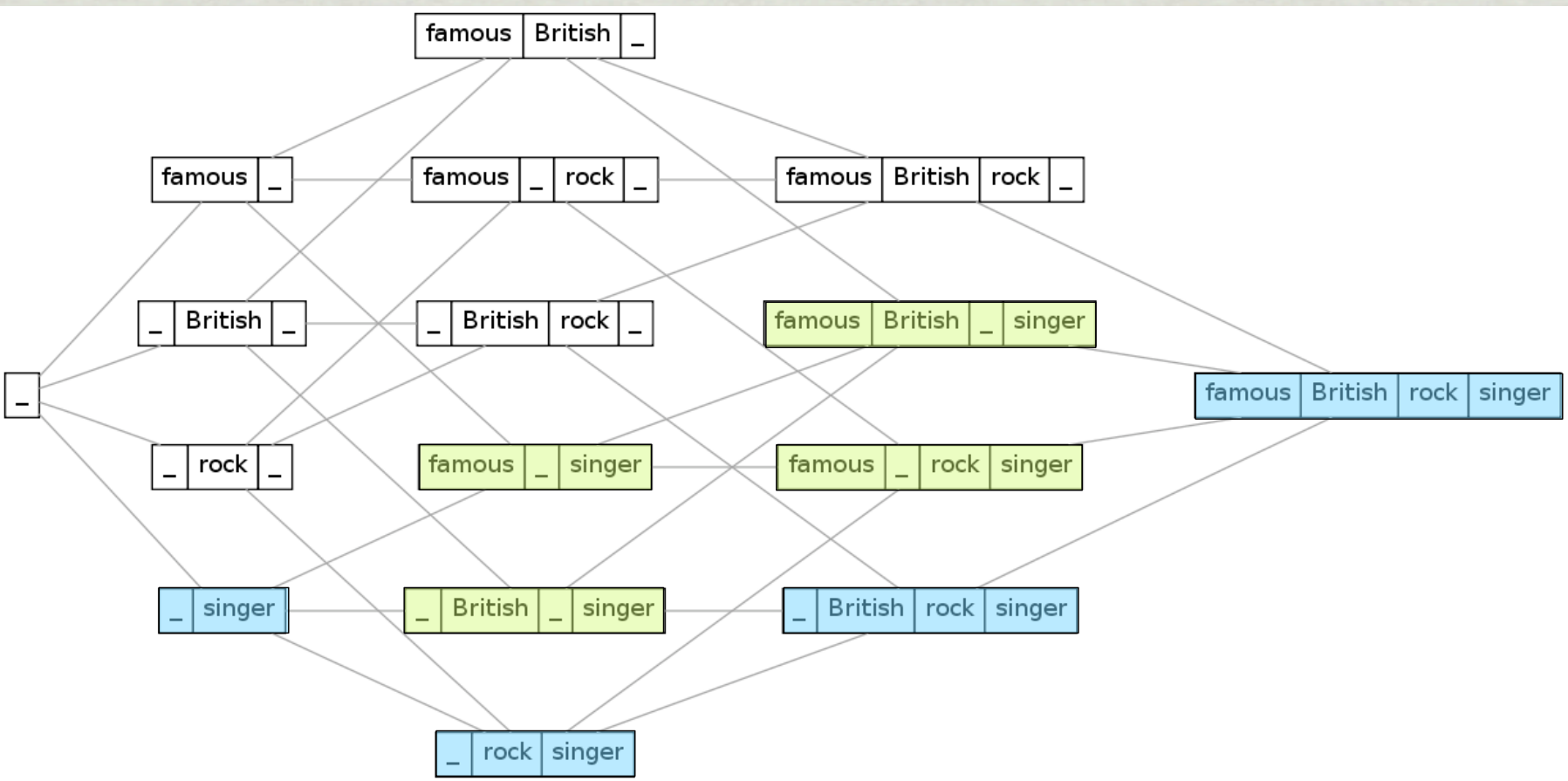
- Given a hypernym h_n,
  1. we automatically generated hypernym path $H(h_n) = (h_1, h_2, ..., h_n)$ (say, using POS information of h_n).
  2. then manually checked if h_i is a valid word or not.
- Remark
  - We worked only on Japanese examples, though we will present English examples in this talk for expository purposes.

NiCT National Institute of Information and Communications Technology

MASTAR

GTA, Simplified or Not

We performed simplified GTA that needs to be distinguished from full GTA where both blue and green units are identified.



generated `rubyplb` available at http:/www.kotonoba.net/pattern

Sample of Simplified GTA

	hypernym1	hypernym2	hypernym3	hypernym4	hyponym
1	人 (person)	料理人 (cook)	フランス料理人 (French cook)		坂井宏行 (Sakai, Hiroyuki)
2	品* (item)	製品 (product)	ドイツの製品 (product of Germany)		ペリーローダンRPG (Perry Rhodan RPG)
3	品* (item)	用品 (items for ...)	園芸用品 (gardening supply)		ワイパアゾル (Wiper-sol)
4	品* (item)	作品 ((piece of) work)	題材にした作品 ((piece of) work on ...)	吸血鬼を題材にした作品 ((piece of) work on vampries)	Black Blood Brothers
5	家* (agent)	運動家 (activist)	フェミニズム運動家 (feminism activist)		テロワーニュ・ド・メリクール (Théroigne de Méricourt)
6	家* (family)	五家 ((major) five families)	禅宗五家 ((major) five schools of Zen)	中国禅宗五家 ((major) five schools of Chinese Zen)	臨済宗 (Rinzai school of Zen)
7	手* (agent)	騎手 (jockey)	イギリスの騎手 (British jockey)		キーレン・ファロン (Kieren Fallon)
8	手* (agent)	選手 (player)	野球選手 (Baseball player)	プエルトリコの野球選手 (Baseball player in Puerto Rico)	イバン・クルーズ (Luis Iván Cruz)
9	社* (site of sacred)	神社 (shrine)	市の神社 (shrine of a City)	鎌倉市の神社 (shrine of Kamakura City)	龍口明神社
10	社* (company)	出版社 (publisher)	音楽出版社 (music publisher)		音楽之友社

Units with *, typically at leftmost, are units smaller than words

NiCT National Institute of Information and Communications Technology

MA STAR

GTA in Action
sample English examples

- GTA is not a trivial task. It needs to deal with cases like the following

	Type						Type					
1	L	former	member	of	Pink	Floyd	L	famous	product	of	West	Germany
2	G		member	of	Pink	Floyd	G		product	of	West	Germany
3	B		member	of		Floyd	G		product	of		Germany
4	L		member				L		product			

  - Lebels: (i) *G* for proper, saturated, (ii) *L* for proper, unsaturated, and (iii) *B* for improper

- GTA requires adequate analysis of modification structure.

Challenges in GTA

A. distinguishing proper phrases from improper phrases.

- Set of of "proper' phrases is conventionally constrained and is far smaller than combinatorially possible set.
- Also, A is affected by *semantically unsaturated nouns (SUNs)* (Kuroda et al. 2009; Nishiyama 1990, 2003), which are a superclass of relational nouns (de Bruin and Scha 1988).

B. If A is satisfied, we need to deal with conventional (often idiomatic) expressions without transparent, compositional semantics.

Challenges in GTA: Noise

- 製品 (product of ...) is a proper word/term in Japanese.
  - 鉄製品 (product from iron), アメリカ製品 (product of America)
- But *用品 (items for ...) is not (or rather hardly so).
  - 日用品 (items for daily use), 車用品 (items for car), 園芸用品 (items for gardening), cf. 旅行の用品店 (shop for travel gear)
- No really semantic account for such differences.

Challenges in GTA: SUNs

- Alleged semantically unsaturated nouns include:
  - *player* in GAME, *winner* of COMPETITION, *disciple* of MASTER, *brother* of PERSON, *father* of PERSON, *father* of PRODUCT, IDEA (metaphorical)
  - *member* of {GROUP, TEAM, ...}, *alumini* of SCHOOL
  - *album* by ARTIST, *track* of ALBUM, *product* of {COMPANY, COUNTRY, ...},
  - *technique(s)* in PRACTICE
- Importantly, frequent hypernyms tend to be SUNs.

Random Sample of Hypernym-Hyponym Pairs from English Wikipedia (Oh et al. 2009)

	Hypernym	Hyponym	SVM Score
1	albums	Time To Say Goodbye/Timeless	1.34114
2	albums	No Fish Shop Parking	1.09981
3	all judges	Winder Laird Henry	0.895937
4	alumni	Mike Corbett	1.34561
5	awards	Artios nominated for Best Casting for TV	0.805839
6	birds of Spain	Recurvirostridae	0.838847
7	forensic anthropologists	Turhon A. Murad	0.821139
8	highways numbered 399	Quebec Route 399	0.904606
9	mayors of Amsterdam	Pieter Claesz van Neck	1.15704
10	national historic sites of Canada	Masonic Memorial Temple	1.05046
11	Newfoundland and Labrador parks	Topsail Beach	1.17714
12	Public Health and Health Services Division	Centre for Prevention and Health Services Research	0.971706
13	recordings	Stop	0.838389
14	track	Bad Obsession	1.14978
15	track	Before I Leap	1.18942
16	track	On My Pillow	0.942252
17	typical antbirds	Chapman's Antshrike Thamnophilus zarumae	1.2905
18	winners	Evelyn Waugh	1.03602
19	works by heads of state or government	The Downing Street Years	1.14225
20	writers and publications	Hugh J. Schonfield	0.958676

We are hardly happy with pairs with unsaturated hypernyms (in orange) that do not serve as good sortal.

NiCT National Institute of Information and Communications Technology

MASTAR

Random Sample of Hypernym-Hyponym Pairs from English Wikipedia (Oh et al. 2009)

	Hypernym	Hyponym	SVM Score
1	albums	Time To Say Goodbye/Timeless	1.34114
2	albums	No Fish Shop Parking	1.09981
3	all judges	Winder Laird Henry	0.895937
4	alumni	Mike Corbett	1.34561
5	awards	Artios nominated for Best Casting for TV	0.805839
6	birds of Spain	Recurvirostridae	0.838847
7	forensic anthropologists	Turhon A. Murad	0.821139
8	highways numbered 399	Quebec Route 399	0.904606
9	mayors of Amsterdam	Pieter Claesz van Neck	1.15704
10	national historic sites of Canada	Masonic Memorial Temple	1.05046
11	Newfoundland and Labrador parks	Topsail Beach	1.17714
12	Public Health and Health Services Division	Centre for Prevention and Health Services Research	0.971706
13	recordings	Stop	0.838389
14	track	Bad Obsession	1.14978
15	track	Before I Leap	1.18942
16	track	On My Pillow	0.942252
17	typical antbirds	Chapman's Antshrike Thamnophilus zarumae	1.2905
18	winners	Evelyn Waugh	1.03602
19	works by heads of state or government	The Downing Street Years	1.14225
20	writers and publications	Hugh J. Schonfield	0.958676

We are hardly happy with pairs with unsaturated hypernyms (in orange) that do not serve as good sortal.

NiCT National Institute of Information and Communications Technology

MA STAR

Challenges in GTA: Idioms

- Are the following abstractions valid or not?
  - secret *weapon* ?*⇒ *weapon*
  - world *heritage* ?⇒ *heritage*
  - electric *piano* ??⇒ *piano*
    - (metaphorical) sense extension is messy as usual.
- A high proportion of compound nouns can be idiomatic, but there is no effective method to detect them automatically.

Result

- We decided to perform GTA manually for all hypernyms.
  - This resulted in ~67,000 hierarchies (released through ALAGIN (Advanced LAnGuage INformation) (http://www.alagin.jp/)
- In theory, GTA *can* be automatized, but we need
  - either manual construction of transformation rules or preparation of training data with good quality for machine learning.
- Our results can be used for either purpose.

NiCT National Institute of Information and Communications Technology

MASTAR

Linking Wikipedia-derived data to the Japanese WordNet

Problem

- Only 8% of the "raw" hypernyms H_0 of the original pairs appear in the lemmas of Japanese WordNet (JWN) (Bond et al. 2008, 2009).
- Reason
  1. H_0 contained a large number of complex phrases with modifiers (~70%).
  2. H_0 contained notational variations (5~10% of the data)
- We can expect this is automatically solved by GTA.

NiCT National Institute of Information and Communications Technology

MASTAR

Remark on Japanese WN

- For now, JWN is just a Japanese translation of Princeton WordNet 3.0.
- It is not a WordNet for Japanese built from scratch and shows a number of troubles with:
  - lexical concepts particular to Japanese, and
  - lexical concepts that are “alian” to Japanese conceptualization
  - part-of-speech mismatch issue, especially with adjectival nouns.

NiCT National Institute of Information and Communications Technology

MASTAR

Effect of GTA

depth	# of hyponyms covered	coverage ratio	# of hypernym types
1	64,412	0.9592	3,272
2	24,554	0.3657	2,447
3	2804	0.0418	465
4	53	0.0008	30

NiCT National Institute of Information and Communications Technology

MASTAR

Remaining issues

- Links from Wikipedia-derived data to JWN lemmas do not undergo sense disambiguation. This was left for future work.
- Yamada et al. (to appear) proposed a method to do it automatically.
  - In a nutshell, sense disambiguation of hypernym *h* can be achieved by "voting" from contextually similar hyponyms s_1, s_2, ..., s_n, selected using similarity data developed by Kazama et al (2008).
  - Informal evaluation results in ~90% accuracy.

Conclusions

- After GTA, 95% of Wikipedia-derived hypernyms are linked to lemmas of JWN (5% to 95% increase).
- This suggests that usefulness of Wikipedia-derived data is limited unless automatic GTA with high precision is implemented.
- In other words, Wikipedia-derived data as it stands cannot dispense with lexical resources like WN.
  - The two kinds of data are best understood as complementary to each other.

NiCT National Institute of Information and Communications Technology

MASTAR

Acknowledgements

- We thank
  - Jong-Hoon Oh (NICT) for giving us hypernym-hyponym pairs acquired from the English Wikipedia.
  - Ichiro Yamada (NICT) for informing us of a method for automatic synset disambiguation.

NiCT National Institute of Information and Communications Technology

MASTAR

Thank you!


NiCT
National Institute of
Information and
Communications
Technology


MA STAR

References

- Bond, F., H. Isahara, K. Kanzaki, and K. Uchimoto (2008). Boot-strapping a WordNet using multiple existing WordNets. In *Proc. of the 6th International Conference on LREC-2008*.
- Bond, F., H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and Kyoko Kanzaki (2009). Enhancing the Japanese WordNet. In *Proc. of the 7th Workshop on Asian Language Resources*, pp. 1–8.
- de Bruin, J. and Scha, R. (1988). The interpreation of relational nouns. In *Proc. of the 26th Annual Meeting of the ACL*, pp. 25–32.
- Fellbaum, C., ed. (1988). *WordNet: An Electric Lexical Database*. MIT Press.
- Levy, P. (1997). *Collective Intelligence*. Basic Books.
- Kazama, J., S. de Sager, K. Torisawa, and M. Murata (2009). Constructing large-scale database of similar nouns using probabilistic clustering of dependency relations. In *Proc. of the 15th Annual Meeting of NLP Association*, pp. 84–87. (In Japanese).
- Kuroda, K., M. Murata and K. Torisawa (2009). When nouns need (co-)arguments: A study of semantically unsaturated nouns. In *Proc. of The 5th International Workshop on Generative Approaches to the Lexicon, 2009, Pisa, Italy*, pp. 193–200.
- Nishiyama, Y. (1990). On the “Kakiryori ha Hiroshima ga honba da” construction: Saturated and unsaturated noun phrases [in Japanese]. In *Proc. of the Institute of Language and Culture, Keio University*, 22, pp. 169–188.
- Nishiyama, Y. (2003) *Semantics and Pragmatics of Noun Phrases in Japanese: Referential and Nonreferential Nouns* [in Japanese]. Hitsuji Publishing.
- Sumida, A. and N. Yoshinaga and K. Torisawa (2008). Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proc. of the 6th International Conference on LREC-2008*.
- Tovey, M., ed. (2008). *Collective Intelligence: Creating a Prosperous World at Peace*. Earth Intelligence Network}
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.
- Yamada, I., et al. (to appear). Adding terms to Japanese WordNet using Wikipedia [in Japanese]. In *Proc. of the 16th Annual Meeting of the NLP Association*.

NiCT National Institute of Information and Communications Technology

MASTAR