A Look inside the Distributionally Similar Terms

Kow Kuroda kuroda@nict.go.jp
Jun'ichi Kazama kazama@nict.go.jp
Kentaro Torisawa torisawa@nict.go.jp
National Institute of Information and Communications Technology (NICT), Japan

Abstract

We analyzed the details of a Web-derived distributional data of Japanese nominal terms with two aims. One aim is to examine if distributionally similar terms can be in fact equated with "semantically similar" terms, and if so to what extent. The other is to investigate into what kind of semantic relations constitute (strongly) distributionally similar terms. Our results show that over 85% of the pairs of the terms derived from the highly similar terms turned out to be semantically similar in some way. The ratio of "classmate," synonymous, hypernym-hyponym, and meronymic relations are about 62%, 17%, 8% and 1% of the classified data, respectively.

1 Introduction

The explosion of online text allows us to enjoy a broad variety of large-scale lexical resources constructed from the texts in the Web in an unsupervised fashion. This line of approach was pioneered by researchers such as Hindle (1990), Grefenstette (1993), Lee (1997) and Lin (1998). At the heart of the approach is a crucial working assumption called "distributional hypothesis," as with Harris (1954). We now see an impressive number of applications in natural language processing (NLP) that benefit from lexical resources directly or indirectly derived from this assumption. It seems that most researchers are reasonably satisfied with the results obtained thus far.

Does this mean, however, that the distributional hypothesis was proved to be valid? Not necessarily: while we have a great deal of confirmative results reported in a variety of research areas, but we would rather say that the hypothesis has never been fully "validated" for two reasons. First, it has yet to be tested under the precise definition of "semantic similarity." Second, it has yet to be tested against results obtained at a truly large scale.

One of serious problems is that we have seen no agreement on what "similar terms" mean and should mean. This paper intends to cast light on this unsolved problem through an investigation into the precise nature of lexical resources constructed under the distributional hypothesis. The crucial question to be asked is, Can distributionally similar terms really be equated with semantically similar terms or not? In our investigation, we sought to recognize what types of semantic relations can be found for pairs of terms with high distributional similarity, and see where the equation of distributional similarity with semantic similarity fails. With this concern, this paper tried to factor out as many components of semantic similarity as possible. The effort of factorization resulted in the 18 classes of semantic (un)relatedness to be explained in §2.3.1. Such factorization is a necessary step for a full validation of the hypothesis. To meet the criterion of testing the hypothesis at a very large scale, we analyzed 300,000 pairs of distributionally similar terms. Details of the data we used are given in §2.2.

This paper is organized as follows. In §2, we present our method and data we used. In §3, we present the results and subsequent analysis. In §4, we address a few remaining problems. In §5, we state tentative conclusions.

2 Method and Data

2.1 Method

The question we need to address is how many subtypes of semantic relation we can identify in the highly similar terms. We examined the question in the following procedure:

(1) a. Select a set of "base" terms B.
b. Use a similarity measure M to construct a list of n terms $T = [t_{i,1}, t_{i,2}, \ldots, t_{i,j}, \ldots, t_{i,n}]$ where $t_{i,j}$ denotes the j-th most similarity term in T against $b_i \in B$. $P(k)$ are pairs of b_i and $t_{i,k}$, i.e., the k-th most similar term to b_i.
c. Human raters classify a portion Q of the pairs in $P(k)$ with reference to a classification guideline prepared for the task.

Note that the selection of base set B can be independent of the selection of T. Note also that T is indexed by terms in B. To encode this, we write: $T[b_i] = [t_{i,1}, t_{i,2}, \ldots, t_{i,j}, \ldots, t_{i,n}]$.

2.2 Data

For T, we used Kazama's nominal term clustering (Kazama and Torisawa, 2008; Kazama et al., 2009). In this data, base set B for T is one million terms defined by the type counts of dependency relations, which is roughly equated with the "frequencies" of the terms. Each base term in B is associated with up to 500 of the most distributionally similar terms. This defines T.

For M, we used the Jensen-Shannon divergence (JS-divergence) base on the probability distributions derived by an EM-based soft clustering (Kazama and Torisawa, 2008). For convenience, some relevant details of the data construction are described in Appendix A, but in a nutshell, we used dependency relations as distributional information. This makes our method comparable to that used in Hindle (1990). The statistics of the distributional data used were as follows: roughly 920 million types of dependency relations[1)] were automatically acquired from a large-scale Japanese Web-corpus called the *Tsubaki* corpus (Shinzato et al., 2008) which consists of roughly 100 million Japanese pages with six billion sentences. After excluding hapax nouns, we had about 33 million types of nouns (in terms of string) and 27 million types of verbs. These nouns were ranked by type count of the two context triples, i.e., (t, p, v) and (n^*, p^*, t). B was determined by selecting the top one million terms with the most variations of context triples.

2.2.1 Sample of $T[b]$

For illustration, we present examples of the Web-derived distributional similar terms. (2) shows the 10 most distributionally similar terms (i.e., $[t_{1070,1}, t_{1070,2}, \ldots, t_{1070,10}]$ in $T(b_{1070})$) where b_{1070} = "ピアノ" (piano) is the 1070-th term in B. Likewise, (3) shows the 10 most distributionally similar terms $[t_{38555,1}, t_{38555,2}, \ldots, t_{38555,10}]$ in $T(b_{38555})$) where b_{38555} = "チャイコフスキー" (Tchaikovsky) is the 38555-th term in B.

(2) 10 most similar to "ピアノ"

1. エレクトーン (Electone; electronic organ) [-0.322]
2. バイオリン (violin) [-0.357]
3. ヴァイオリン (violin) [-0.358]
4. チェロ (cello) [-0.358]
5. トランペット (trumpet) [-0.377]
6. 三味線 (shamisen) [-0.383]
7. サックス (saxophone) [-0.39]
8. オルガン (organ) [-0.392]
9. クラリネット (clarinet) [-0.394]
10. 二胡 (erh hu) (-0.396)

(3) 10 most similar to "チャイコフスキー"

1. ブラームス (Brahms) [-0.152]
2. シューマン (Schumann) [-0.163]
3. メンデルスゾーン (Mendelssohn) [-0.166]
4. ショスタコーヴィチ (Shostakovich) [-0.178]
5. シベリウス (Sibelius) [-0.18]

[1)] The 920 million types come in two kinds of context triples: 590 million types of (t, p, v) and 320 million types of (t, p^*, n^*), where t denotes the target nominal term, p a postposition, v a verb, and n^* a nominal term that follows t and p^*, i.e., "*t-no*" analogue to the English "*of t*."

6. ハイドン (Haydn) [-0.181]
7. ヘンデル (Händel) [-0.181]
8. ラヴェル (Ravel) [-0.182]
9. シューベルト (Schubert) [-0.187]
10. ベートーヴェン (Beethoven) [-0.19]

For construction of $P(k)$, we had the following settings: i) $k = 1, 2$; and ii) for each k, we selected the 150,000 most frequent terms (out of one million terms) with some filtering specified below. Thus, Q was 300,000 pairs whose base terms are roughly the most frequent 150,000 terms in B with filtering and targets are terms $k = 1$ or $k = 2$.

2.2.2 Filtering of terms in B

For filtering, we excluded the terms of B with one of the following properties: a) they are in an invalid form that could have resulted from parse errors; b) they have regular ending (e.g., -こと, -事 [event], -時 [time or when], -もの [thing or person], -物 [thing], -者 [person]). The reason for the second is two-fold. First, it was desirable to reduce the ratio of the class of "classmates with common morpheme," which is explained in §2.3.2, whose dominance turned out to be evident in the preliminary analysis. Second, the semantic property of the terms in this class is relatively predictable from their morphology. That notwithstanding, this filtering might have had an undesirable impact on our results, at least in terms of representativeness. Despite of this, we decided to place priority on collecting more varieties of classes.

The crucial question is, again, whether distributionally similar terms can really be equated with semantically similar terms. Put differently, what kinds of terms can we find in the sets constructed using distributionally similarity? We can confirm the hypothesis if the most of the term pairs are proved to be semantically similar for most sets of terms constructed based on the distributional hypothesis. To do this, however, we need to clarify what constitutes semantic similarity. We will deal with this prerequisite.

2.3 Classification

2.3.1 Factoring out "semantic similarity"

Building on lexicographic works like Fellbaum (1998) and Murphy (2003), we assume that the following are the four major classes of semantic relation that contribute to semantic similarity between two terms:

(4) a. "synonymic" relation (one can substitute for another on an identity basis). Examples are (*Microsoft*, *MS*).

b. "hypernym-hyponym" relation between two terms (one can substitute for another on un underspecification/abstraction basis). Examples are (*guys*, *players*)

c. "meronymic" (part-whole) relation between two terms (one term can be a substitute for another on metonymic basis). Examples are (*bodies*, *players*) [cf. *All the players have strong bodies*]

d. "classmate" relation between two terms, t_1 and t_2, if and only if (i) they are not synonymous and (ii) there is a concrete enough class such that both t_1 and t_2 are instances (or subclasses).[2)] For example, (*China*, *South Korea*) [cf. *(Both) China and South Korea are countries in East Asia*], (*Ford*, *Toyota*) [cf. *(Both) Ford and Toyota are top-selling automotive companies*] and (*tuna*, *cod*) [cf. *(Both) tuna and cod are types of fish that are eaten in the Europe*] are classmates.

For the substitution, the classmate class behaves somewhat differently. In this case, one term cannot substitute for another for a pair of terms. It is hard to find out the context in which pairs like (*China*, *South Korea*), (*Ford*, *Toyota*) and (*tuna*, *cod*) can substitute one another. On the other hand, substitution is more or less possible in the other three types. For example, a synonymic pair of (*MS*, *Microsoft*) can substitute for one another in contexts like *Many people regularly complain*

[2)] The proper definition of classmates is extremely hard to form. The authors are aware of the incompleteness of their definition, but decided not to be overly meticulous.

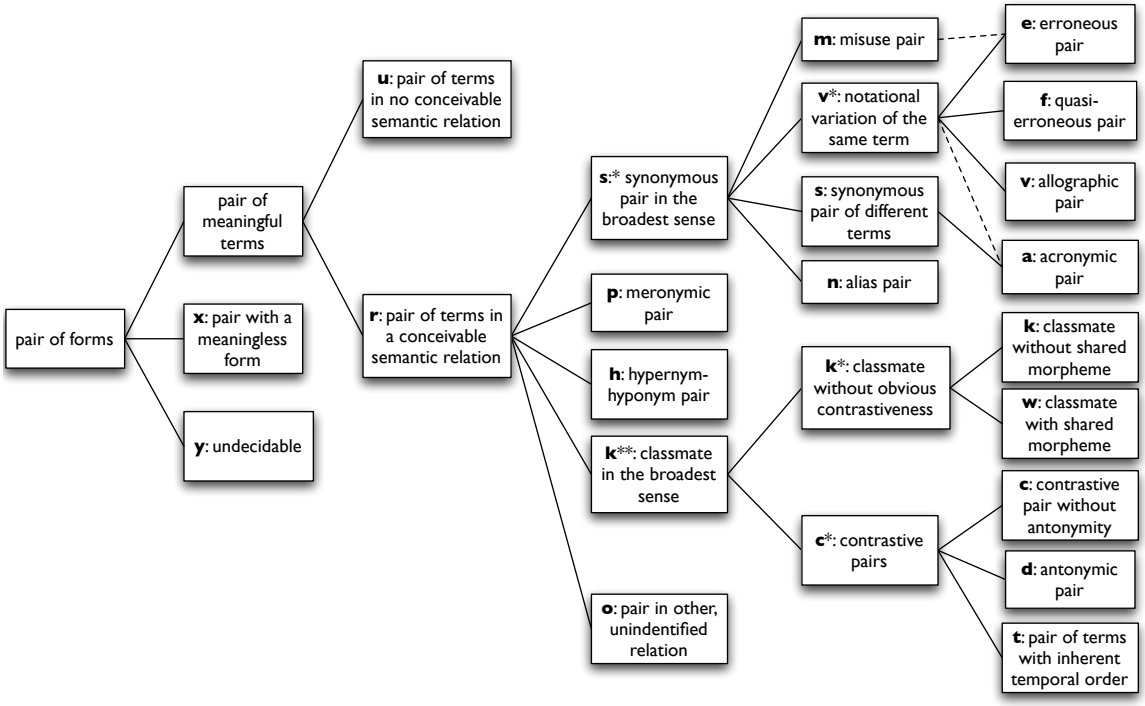

Figure 1: Classification tree for semantic relations used

about products { *i. MS; ii. Microsoft* }. A hypernym-hyponym pair of (*guys*, *players*) can substitute in contexts like *We have dozens of excellent* { *i. guys; ii. players* } *on our team*. A meronymic pair of (*bodies*, *players*) can substitute for each other in contexts like *They had a few of excellent* { *i. bodies; ii. players* } *last year*.

2.3.2 Classification guidelines

The classification guidelines were specified based on a preliminary analysis of 5,000 randomly selected examples. We asked four annotators to perform the task. The guidelines were finalized after several revisions. This revision process resulted in a hierarchy of binary semantic relations as illustrated in Figure 1, which subsumes 18 types as specified in (5). The essential division is made at the fourth level where we have **s*** (pairs of synonyms in the broadest sense) with two subtypes, **p** (pairs of terms in the "part-whole" relation), **h** (pairs of terms in the "hypernym-hyponym" relation), **k**** (pairs of terms in the "classmate" relation), and **o** (pairs of terms in any other relation). Note that this system includes the four major types described in (4). The following are some example pairs of Japanese terms with or without English translations:

(5) **s**: synonymic pairs (subtype of **s***) in which the pair designates the same entity, property, or relation. Examples are: (根元, 株元) [both mean *root*], (サポート会員, 協力会員) [(*supporting member*, *cooperating member*)], (呼び出し元, 親プロセス) [(*invoker of the process*, *parent process*)], (ベンチャービジネス, ベンチャー) [(*venture business*, *venture*)], (相手投手, 相手ピッチャー) [(*opposing hurler*, *opposing pitcher*)], (病歴, 既往歴) [(*medical history*, *anamneses*)],

n: alias pairs (subtype of **s***) in which one term of the pair is the "alias" of the other term. Examples are (*Steve Jobs*, *founder of Apple, Inc.*), (*Barak Obama*, *US President*), (ノグチ, イサム・ノグチ), (侑一郎, うにっ子)

a: acronymic pair (subtype of **s***) in which one term of the pair is the acronym of of the other term. Examples are: (*DEC*, *Digital Equipment*), (*IBM*, *International Business Machine*) (Microsoft 社, MS 社), (難関大, 難関大学), (配置転換, 配転),

v: allographic pairs (subtype of **s***) in which the pair is the pair of two forms of the same term. Examples are: (*convention centre*, *convention center*), (*colour terms*, *color terms*), (乙女ゲーム, 乙女ゲー), (アカスリ, あかすり), (コンピュータシステム, コンピューターシステム), (廻り, 回り), (Ｓｏｌｏ, ｓｏｌｏ), (かっこ, 括弧), (消化器癌, 消化器がん), (坐薬, 座薬), (踏みつけ, 踏み付け)

h: hypernym-hyponym pair in which one term of the pair designates the "class" of the other term. Examples (order is irrelevant) are: (thesaurus, Roget's), (検索ツール, 検索ソフト) [(*search tool*, *search software*)], (失業対策, 雇用対策) [(*unemployment measures*, *employment measures*)], (景況, 雇用情勢) [(*business conditions*, *employment conditions*)], (フェスティバル, 音楽祭) [(*festival*, *music festival*)], (検査薬, 妊娠検査薬) [(*test agent*, *pregnancy test*)], (シンビジウム, 洋ラン) [(*cymbidium*, *orchid*)], (企業ロゴ, ロゴマーク) [(*company logo*, *logo*)], (神秘体験, 臨死体験) [(*mystical experiences*, *near-death experiences*)]

p: meronymic pair in which one term of the pair designates the "part" of the other term. Examples (order is irrelevant) are: (ちきゅう, うみ) [(*earth*, *sea*)], (確約, 了解) [(*affirmation*, *admission*)], (知見, 研究成果) [(*findings*, *research progress*)], (ソーラーサーキット, 外断熱工法) [(*solar circuit system*, *exterior thermal insulation method*)], (プロバンス, 南仏) [(*Provence*, *South France*)],

k: classmates not obviously contrastive without common morpheme (subtype of **k***). Examples are: (自分磨き, 体力作り) [(*self-culture*, *training*)], (所属機関, 部局) [(*sub-organs*, *services*)], (トンパ文字, ヒエログリフ) [(*Dongba alphabets*, *hieroglyphs*)], (*Tom*, *Jerry*)

w: classmates not obviously contrastive with common morpheme (subtype of **k***). Examples are: (ガス設備, 電気設備) [(*gas facilities*, *electric facilities*)], (他社製品, 本製品) [(*products of other company*, *aforementioned products*)], (系列局, 地方局) [(*affiliate station*, *local station*)], (新潟市, 和歌山市) [(*Niigata City*, *Wakayama City*)], (シナイ半島, マレー半島) [(*Sinai Peninsula*, *Malay Peninsula*)],

c: contrastive pairs without antonymity (subtype of **c***). Examples are: (ロマン主義, 自然主義) [(*romanticism*, *naturalism*)], (携帯電話ユーザー, インターネットユーザー) [(*mobile user*, *internet user*)], (海外版, PS2 版), [(*bootleg edition*, PS2 *edition*)]

d: antonymic pairs = contrastive pairs with antonymity (subtype of **c***). Examples are: (接着, 分解) [(*bonding*, *disintegration*)], (砂利道, 舗装路) [(*gravel road*, *pavement*)], (西壁, 東壁) [(*west walls*, *east walls*)], (娘夫婦, 息子夫婦) [(*daughter and son-in-law*, *son and daughter-in-law*)], (外税, 内税) [(*tax-exclusive prices*, *tax-inclusive prices*)], (リアブレーキ, フロントブレーキ) [(*front brake*, *rear brake*)], (タッグマッチ, シングルマッチ) [(*tag-team match*, *solo match*)], (乾拭き, 水拭き) [(*wiping with dry materials*, *wiping with wet materials*)], (ノースリーブ, 長袖) [(*sleeveless*, *long-sleeved*)]

t: pairs with inherent temporal order (subtype of **c***). Examples are: (稲刈り, 田植え) [(*harvesting of rice*, *planting of rice*)], (ご到着日, ご出発日) [(*day of departure*, *day of arrival*)], (進路決定, 進路選択) [(*career decision*, *career selection*)], (居眠り, 夜更かし) [(*catnap*, *stay up*)], (密猟, 密輸) [(*poaching*, *con-*

traband trade)], (投降, 出兵) [(*surrender*, *dispatch of troops*)], (二回生, 三回生) [(*2nd-year student*, *3rd-year student*)]

e: erroneous pairs are pairs in which one term of the pair seems to suffer from character-level input errors, i.e. "mistypes." Examples are: (筋線維, 筋繊維), (発砲スチロール, 発泡スチロール), (太宰府, 大宰府)

f: quasi-erroneous pair is a pair of terms with status somewhat between **v** and **e**. Examples (order is irrelevant) are: (スポイト, スポイド) [(*supoi***t***o*, *supoi***d***o*)], (ゴルフバッグ, ゴルフバック) [(*goru-fuba***gg***u*, *gorufuga***kk***u*)], (ビックバン, ビッグバン) [(*bi***gg***u ban*, *bi***kk***u ban*)],

m: misuse pairs in which one term of the pair seems to suffer from "mistake" or "bad memory" of a word (**e** is caused by mistypes but **m** is not). Examples (order is irrelevant) are: (氷漬け, 氷付け), (積み下ろし, 積み降ろし), (開講, 開校), (恋愛観, 恋愛感), (平行, 並行)

o: pairs in other unidentified relation in which the pair is in some semantic relation other than **s***, **k****, **p**, **h**, and **u**. Examples are: (下心, 独占欲) [(*ulterior motives*, *possessive feeling*)], (理論的背景, 基本的概念) [(*theoretical background*, *basic concepts*)], (アレクサンドリア, シラクサ) [(*Alexandria*, *Siracusa*)],

u: unrelated pairs in which the pair is in no promptly conceivable semantic relation. Examples are: (非接触, 高分解能) [(*noncontact*, *high resolution*)], (模倣, 拡大解釈) [(*imitation*, *overinterpretation*)],

x: nonsensical pairs in which either of the pair is not a proper term of Japanese. (but it can be a proper name with very low familiarity). Examples are: (わったん, まる赤), (セルディ, 瀬璃), (チル, エルダ), (ウーナ, 香瑩), (ma, ジョージア)

y: unclassifiable under the allowed time limit.[3] Examples are: (場所網, 無規準ゲーム), (fj, スラド), (反力, 断力),

Note that some relation types are symmetric and others are asymmetric: **a**, **n**, **h**, **p**, and **t** (and **e**, **f**, and **m**, too) are asymmetric types. This means that the order of the pair is relevant, but it was not taken into account during classification. Annotators were asked to ignore the direction of pairs in the classification task. In the finalization, we need to reclassify these to get them in the right order.

2.3.3 Notes on implicational relations

The overall implicational relation in the hierarchy in Figure 1 is the following:

(6)
a. **s**, **k****, **p**, **h**, and **o** are supposed to be mutually exclusive, but the distinction is sometimes obscure.[4]
b. **k**** has two subtypes: **k*** and **c***.
c. **k** and **w** are two subtypes **k***.
d. **c**, **d** and **t** three subtypes of **c***.

To resolve the issue of ambiguity, priority was set among the labels so that **e**, $\mathbf{f} < \mathbf{v} < \mathbf{a} < \mathbf{n} < \mathbf{p} < \mathbf{h} < \mathbf{s} < \mathbf{t} < \mathbf{d} < \mathbf{c} < \mathbf{w} < \mathbf{k} < \mathbf{m} < \mathbf{o} < \mathbf{u} < \mathbf{x} < \mathbf{y}$, where the left label is more preferred over the right. This guarantees preservation of the implicational relationship among labels.

2.3.4 Notes on quality of classification

We would like to add a remark on the quality. After a quick overview, we reclassified **o** and **w**, because the first run of the final task ultimately produced a resource of unsatisfactory quality.

Another note on inter-annotator agreement: originally, the classification task was designed and run as a part of a large-scale language resource development. Due to its overwhelming size, we tried to make our development as efficient as possible. In the final phase, we asked

[3] We did not ask annotators to check for unknown terms.

[4] To see this, consider pairs like (*large bowel*, *bowel*), (*small bowel*, *bowel*). Are they instances of **p** or **h**? The difficulty in the distinction between **h** and **p** becomes harder in Japanese due to the lack of plurality marking: cases like (*Mars*, *heavenly body*) (a case of **h**) and (*Mars*, *heavenly bodies*) (a **p** case) cannot be explicitly distinguished. In fact, the Japanese term 天体 can mean both "heavenly body" (singular) and "heavenly bodies" (plural).

Table 1: Distribution of relation types

rank	count	ratio (%)	cum. (%)	class	label
1	108,149	36.04	36.04	classmates without common morpheme	**k**
2	67,089	22.35	58.39	classmates with common morpheme	**w**
3	26,113	8.70	67.09	synonymic pairs	**s**
4	24,599	8.20	75.29	hypernym-hyponym pairs	**h**
5	20,766	6.92	82.21	allographic pairs	**v**
6	18,950	6.31	88.52	pairs in other "unidentified" relation	**o**
7	12,383	4.13	92.65	unrelated pairs	**u**
8	8,092	2.70	95.34	contrastive pairs without antonymity	**c**
9	3,793	1.26	96.61	pairs with inherent temporal order	**t**
10	3,038	1.01	97.62	antonymic pairs	**d**
11	2,995	1.00	98.62	meronymic pairs	**p**
12	1,855	0.62	99.23	acronymic pairs	**a**
13	725	0.24	99.48	alias pairs	**n**
14	715	0.24	99.71	erroneous pairs	**e**
15	397	0.13	99.85	misuse pairs	**m**
16	250	0.08	99.93	nonsensical pairs	**x**
17	180	0.06	99.99	quasi-erroneous pairs	**f**
18	33	0.01	100.00	unclassified	**y**

17 annotators to classify the data with no overlap. Ultimately we obtained results that deserve a detailed report. This history, however, brought us to an undesirable situation: no inter-annotator agreement is calculable because there was no overlap in the task. This is why no inter-rater agreement data is now available.

3 Results

Table 1 summarizes the distribution of relation types with their respective ranks and proportions. The statistics suggests that classes of **e**, **f**, **m**, **x**, and **y** can be ignored without risk.

3.1 Observations

We noticed the following. Firstly, the largest class is the class of classmates, narrowly defined or broadly defined. The narrow definition of the classmates is the conjunction of **k** and **w**, which makes 58.39%. The broader definition of classmates, **k****, is the union of **k**, **w**, **c**, **d** and **t**, which makes 62.10%. This confirms the distributional hypothesis.

The second largest class is the narrowly defined synonymous pairs **s**. This is 8.7% of the total, but the general class of synonymic pairs, **s*** as the union of **s**, **a**, **n**, **v**, **e**, **f**, and **m**, makes 16.91%. This comes next to **h** and **w**. Notice also that the union of **k**** and **s*** makes 79.01%.

The third largest is the class of terms in hypernym-hyponym relations. This is 8.20% of the total. We are not sure if this is large or small.

These results look reasonable and can be seen as validation of the distributional hypothesis. But there is something uncomfortable about the the fourth and fifth largest classes, pairs in "other" relation and "unrelated" pairs, which make 6.31% and 4.13% of the total, respectively. Admittedly, 6.31% are 4.13% are not very large numbers, but it does not guarantee that we can ignore them safely. We need a closer examination of these classes and return to this in §4.

3.2 Note on allography in Japanese

There are some additional notes: the rate of allographic pairs [**v**] (6.92%) is rather high.[5)] We suspect that this ratio is considerably higher than the similar results that are to be expected in other

[5)] Admittedly, 6.92% is not a large number in an absolute value, but it is quite large for the rate of allographic pairs.

languages. In fact, the range of notational variations in Japanese texts is notoriously large. Many researchers in Japanese NLP became to be aware of this, by experience, and claim that this is one of the causes of Japanese NLP being less efficient than NLP in other (typically "segmented") languages. Our result revealed only the allography ratio in nominal terms. It is not clear to what extent this result is applied to the notional variations on predicates, but it is unlikely that predicates have a lesser degree of notational variation than nominals. At the least, informal analysis suggests that the ratio of allography is more frequent and has more severe impacts in predicates than in nominals. So, it is very unlikely that we had a unreasonably high rate of allography in our data.

3.3 Summary of the results

Overall, we can say that the distributional hypothesis was to a great extent positively confirmed to a large extent. Classes of classmates and synonymous pairs are dominant. If the side effects of filtering described in §2.2.2 are ignored, nearly 88% (all but **o**, **u**, **m**, **x**, and **y**) of the pairs in the data turned out to be "semantically similar" in the sense they are classified into one of the regular semantic relations defined in (5). While the status of the inclusion of hypernym-hyponym pairs in classes of semantically similar terms could be controversial, this result cannot be seen as negative.

One aspect somewhat unclear in the results we obtained, however, is that highly similar terms in our data contain such a number of pairs in unidentifiable relation. We will discuss this in more detail in the following section.

4 Discussion

4.1 Limits induced by parameters

Our results have certain limits. We specify those here.

First, our results are based on the case of $k = 1, 2$ for $P(k)$. This may be too small and it is rather likely that we did not acquire results with enough representativeness. For more complete results, we need to compare the present results under larger k, say $k = 4, 8, 16, \ldots$. We did not do this, but we have a comparable result in one of the preliminary studies. In the preparation stage, we classified samples of pairs whose base term is at frequency ranks 13–172, 798–1,422 and 12,673–15,172 where k = 1, 2, 3, …, 9, 10.[6] Table 2 shows the ratios of relation types for this sample (k = 1, 2, 4, 8, 10).

Table 2: Similarity rank = 1, 2, 4, 8, 10

rank	1	2	4	8	10
v	18.13	10.48	3.92	2.51	1.04
o	17.08	21.24	26.93	28.24	29.56
w	13.65	13.33	14.30	12.19	12.75
s	11.74	9.14	7.05	4.64	4.06
u	11.07	16.48	17.63	20.79	20.87
h	10.50	10.29	11.17	12.96	10.20
k	7.82	8.38	7.84	7.74	8.22
d	2.58	2.00	1.57	1.16	0.85
p	2.00	1.14	1.08	1.35	1.79
c	1.43	1.05	1.27	1.35	1.89
a	1.05	1.33	0.88	0.39	0.57
x	1.05	1.14	1.27	1.64	2.08
t	0.29	0.19	0.20	0.39	0.47
f	0.10	0.10	0.00	0.10	0.09
m	0.00	0.10	0.20	0.00	0.19
#item	1,048	1,050	1,021	1,034	1,059

From Table 2, we notice that: as similarity rank decreases, (i) the ratios of **v**, **s**, **a**, and **d** decrease monotonically, and the ratios of **v** and **s** decrease drastically; (ii) the ratios of **o**, **u**, and **x** increases monotonically, and the ratio of **o** and **u** increases considerably; and while (iii) the ratios of **h**, **k**, **p**, **w**, **m**, and **f** seem to be constant. But it is likely that the ratios of **h**, **k**, **p**, **w**, **m**, and **f** change at larger k, say 128, 256.

Overall, however, this suggests that the difference in similarity rank has the greatest impact on **s*** (recall that **s** and **v** are subtypes of **s***), **o**, and **u**, but not so much on others. Two tendencies can be stated: first, terms at lower similarity ranks become less synonymous. Second,

[6] The frequency/rank in B was measured in terms of the count of types of dependency relation.

the relationships among terms at lower similarity ranks become more obscure. Both are quite understandable.

There are, however, two caveats concerning the data in Table 2, however. First, the 15 labels used in this preliminary task are a subset of the 18 labels used in the final task. Second, the definitions of some labels are not completely the same even if the same labels are used (this is why we have this great of a ratio of **o** in Table 2. We must admit, therefore, that no direct comparison is possible between the data in Tables 1 and 2.

Second, it is not clear if we made the best choices for clustering algorithm and distributional data. For the issue of algorithm, there are too many clustering algorithms and it is hard to reasonably select candidates for comparison. We do, however, plan to extend our evaluation method to other clustering algorithms. Currently, one of such options is Bayesian clustering. We are planning to perform some comparisons.

For the issue of what kind of distributional information to use, many kinds of distributional data other than dependency relation are available. For example, simple co-occurrences within a "window" are a viable option. With a lack of comparison, however, we cannot tell at the present what will come about if another kind of distributional data was used in the same clustering algorithm.

4.2 Possible overestimation of hypernyms

A closer look suggests that the ratio of hypernym-hyponym pairs was somewhat overestimated. This is due to the algorithm used in our data construction. It was often the case that head nouns were extracted as bare nouns from complex, much longer noun phrases, sometimes due to the extraction algorithms or parse errors. This resulted in accidental removal of modifiers being attached to head nouns in their original uses. We have not yet checked how often this was the case. We are aware that this could have resulted in the overestimation of the ratio of hypernymic relations in our data.

4.3 Remaining issues

As stated, the fourth largest class, roughly 6.31% of the total, is that of the pairs in the "other" unidentified relation [**o**]. In our setting, "other" means that it is in none among the synonymous, classmate, part-whole or hypernym-hyponym relation. A closer look into some examples of **o** suggest that they are pairs of terms with extremely vague association or contrast.

Admittedly, 6.31% is not a large number, but its ratio is comparable with that of the allographic pairs [**v**], 6.92%. We have no explanation why we have this much of an unindenfiable kind of semantic relation distinguished from unrelated pairs [**u**]. All we can say now is that we need further investigation into it.

u is not as large as **o**, but it has a status similar to **o**. We need to know why this much amount of this kind of pairs. A possible answer would be that they are caused by parse errors, directly or indirectly.

5 Conclusion

We analyzed the details of the Japanese nominal terms automatically constructed under the "distributional hypothesis," as in Harris (1954). We had two aims. One aim was to examine to see if what we acquire under the hypothesis is exactly what we expect, i.e., if distributional similarity can be equated with semantic similarity. The other aim was to see what kind of semantic relations comprise a class of distributionally similar terms.

For the first aim, we obtained a positive result: nearly 88% of the pairs in the data turned out to be semantically similar under the 18 criteria defined in (5), which include hypernym-hyponym, meronymic, contrastive, and synonymic relations. Though some term pairs we evaluated were among none of these relations, the ratio of **o** and **u** in sum is about 14% and within the acceptable range.

For the second aim, our result revealed that the ratio of the classmates, synonymous, relation, hypernym-hyponym, and meronymic relations are respectively about 62%, 17%, 8% and 1% of the classified data.

Overall, these results suggest that automatic acquisition of terms under the distributional hypothesis give us reasonable results.

A Clustering of one million nominals

This appendix provides some details on how the clustering of one million nominal terms was performed.

To determine the similarity metric of a pair of nominal terms (t_1, t_2), Kazama et al. (2009) used the Jensen-Shannon divergence (JS-divergence) $D_{JS}(p||q) = \frac{1}{2}D(p||M) + \frac{1}{2}D(q||M)$, where p and q are probability distributions, and $D = \sum_i p(i)\log\frac{p(i)}{q(i)}$ (Kullback-Leibler divergence, or KL-divergence) of p and q, and $M = \frac{1}{2}(p+q)$. We obtained p and q in the following way.

Instead of using raw distribution, Kazama et al. (2009) applied smoothing using EM algorithm (Rooth et al., 1999; Torisawa, 2001). In Torisawa's model (2001), the probability of the occurrence of the dependency relation $\langle v, r, n\rangle$ is defined as:

$$P(\langle v,r,t\rangle) =_{\text{def}} \sum_{a\in A} P(\langle v,r\rangle|a)\,P(t|a)\,P(a),$$

where a denotes a hidden class of $\langle v, r\rangle$ and term t. In this equation, the probabilities $P(\langle v,r\rangle|a)$, $P(t|a)$, and $P(a)$ cannot be calculated directly because class a is not observed in a given dependency data. The EM-based clustering method estimates these probabilities using a given corpus. In the E-step, the probability $P(a|\langle v,r\rangle)$ is calculated. In the M-step, the probabilities $P(\langle v,r\rangle|a)$, $P(t|a)$, and $P(a)$ are updated until the likelihood is improved using the results of the E-step. From the results of this EM-based clustering method, we can obtain the probabilities $P(\langle v,r\rangle|a)$, $P(t|a)$, and $P(a)$ for each $\langle v,r\rangle$, t, and a. Then, $P(a|t)$ is calculated by the following equation:

$$P(a|t) = \frac{P(t|a)\,P(a)}{\sum_{a\in A} P(t|a)\,P(a)}.$$

The distributional similarity between t_1 and t_2 was calculated by the JS divergence between $P(a|t_1)$ and $P(a|t_2)$.

References

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Grefenstette, G. 1993. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *In Making Sense of Words: The 9th Annual Conference of the UW Centre for the New OED and Text Research*.

Harris, Z. S. 1954. Distributional structure. *Word*, 10(2-3):146–162. Reprinted in Fodor, J. A and Katz, J. J. (eds.), *Readings in the Philosophy of Language*, pp. 33–49. Englewood Cliffs, NJ: Prentice-Hall.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pp. 268–275, Pittsburgh, PA.

Kazama, J. and K. Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-2008: HLT*, pp. 407–415.

Kazama, J., S. De Saeger, K. Torisawa, and M. Murata. 2009. Generating a large-scale analogy list using a probabilistic clustering based on noun-verb dependency profiles. In *Proceedings of the 15th Annual Meeting of the Association for Natural Language Processing*. [in Japanese].

Lee, L. 1997. *Similarity-Based Approaches to Natural Language Processing*. Unpublished Ph.D. thesis, Harvard University.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL-98, Montreal, Canda*, pages 768–774.

Murphy, M. L. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press, Cambridge, UK.

Rooth, M., S. Riezler, D. Presher, G. Carroll, and F. Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111.

Shinzato, K., T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi. 2008. *TSUBAKI*: An open search engine infrastructure for developing new information access. In *Proceedings of IJCNLP 2008*.

Torisawa, K. 2001. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 211–218.