

Automatic Translation Of WordNet In MLSN

Darren Cook
darren@dcook.org

Introduction

MLSN [1] is an open source project, started in late 2005, to create a semantic network that includes both dictionary words and culture words, and that is multilingual, covering all the world's major languages. The project is first and foremost a database of words and their relations. But it is also a web-based front-end [2] for people to view and enhance that database. It is also a set of utilities for users to make their own local version enhanced with, for instance, their company's confidential data. The open source license freely allows commercial use in the hope that some percentages of business users will contribute back or sponsor enhancements.

The great hope in making a project open source is that people will flock to the project and contribute. The reality is that the project must be useful in order to attract users, and that only some small percentage of those users will become contributors. The catch here is obvious: without contributors it will never have enough data to be useful.

We tackle this issue by trying to use freely available dictionaries to automatic translate entries. This is very far from trivial but all is not lost as there are many words that do have simple unambiguous translations.

Our starting point is Princeton's WordNet [3] for reasons stated in [1]. Though we have added a number of additional terms to the English WordNet the results presented in this paper are based on a clean version of WordNet 3.0.

The Algorithm

Our algorithm is relatively simple and intuitive. It first divides WordNet words into ones it can match and those it cannot, and then categorizes the ones it can match into three groups, labeled as high, medium and low confidence.

The steps are as follows (where en refers to English, and xx refers to our target language):

1. Take all monosemous nouns from WordNet.
2. Translate each using our en-xx dictionary. Only keep those that have exactly one word in the target language.
3. Translate that one word back into English using our xx-en dictionary. Multiple words are fine. Call this set D.
4. For the original noun get its full synset from WordNet. Call this set W.
5. If D=W call it high confidence. Our dictionaries and WordNet agree perfectly.
If D is a perfect subset of W call it medium confidence. WordNet knows more synonyms but our dictionaries did not contradict anything.
If W is a perfect subset of D call it low confidence. In other words our dictionaries found synonyms which WordNet does not know. WordNet is large so this is suspicious.
If anything else throw it away.
6. Import high and medium confidence results, putting xx1h and xx1m, respectively, in the comment field.

The decision to avoid polysemous words is a painful one, as it excludes many common words. But sticking with monosemous words gives quite reliable results. The number of matches in each category is shown in table 1.

	High	Medium	Low
JA	5,130	7,150	3,417
DE	6,795	10,085	3,540
ZH	3,268	5,138	1,589

More Details

To get the list of nouns from WordNet we use WordNet's index.noun file directly. This has one distinct noun per line, and one of the fields on each line is how many synsets it is found in. We first filter this file to just keep those nouns that are in exactly one synset. In other words we throw away all polysemous words, nouns that are in two or more synsets.

For a Japanese dictionary we use Jim Breen's JMDict, including the People and Places extra dictionaries (the Names dictionary was not used) [4]. For German we use BeoLingus [5]. For Chinese we use CEDICT [6]. All three are open source, BeoLingus is GPL while JMDict and CEDICT use a more liberal license that also allows commercial use. These dictionaries are then merged with dictionaries generated from Wikipedia interwiki links as described in [7].

A second independent dictionary source is very important as it helps push suspicious translations into the low confidence category, therefore improving the quality of the high and medium confidence translations. The interwiki dictionary also helps improve quantity of matches due to its good coverage of words not normally found in standard dictionaries (such as products, companies, movies, books, famous people, places, etc.). Taking Japanese as an example, JMDict alone gives us 5,779 high and medium confidence translations, whereas JMDict+Interwiki increases that number to 12,280.

Example

Here are some samples from the English-Japanese intermediate file. It has three fields: the word from WordNet, the single Japanese word that our dictionaries gave us, and then in curly brackets the list of synonyms from the reverse English-Japanese look-up (this is set D in the earlier algorithm description).

```
apple_computer アップルコンピュータ {apple computer}  
apple_pie アップルパイ {apple pie}  
thorn_apple 朝鮮朝顔[ちょうせんあさがお] {datura stramonium,jimsonweed,thorn apple}
```

13apple_pie0 in WordNet contains just "apple pie". This exactly matches the list of English words from our dictionary lookup. Exact is as good as we can get so we flag this as a high confidence match.

14apple_computer0 in WordNet contains two synonyms: "Apple Computer, Apple". Our dictionary lookup found just one of them. In other words, WordNet knows more synonyms for this word than our dictionary, but they both seem to be talking about the same thing. So we flag this as a medium confidence match.

20thorn_apple0 in WordNet contains just "thorn apple". This is only 1 of 3 terms that our dictionary lookup found, so we have low confidence in this result. In other words it may be that 朝鮮朝顔 is a polysemous word. (In this particular example WordNet is to blame: datura stramonium and jimsonweed are synonyms of thorn apple that it does not list.)

Evaluation

The results of the Japanese translation were evaluated, by the author, by choosing 100 random entries from each of the high, medium and low confidence files. The author is a native speaker of English with advanced but imperfect Japanese skills. Therefore personal knowledge was supplemented with dictionaries (GG5 [8], ALC [9]) not used in the above algorithm, careful study of Wikipedia pages to make sure the full nuance of the Japanese was understood, and internet search if the correct answer was still not clear. In addition the hypernym (and other relations) in WordNet were checked to confirm the exact nuance of the English word.

The word lists used in evaluation are included as appendices to this paper, with discussion of troublesome entries appended to some lines. Wrong entries are marked with asterisks, and

problematic entries are marked with question marks.

The results were as follows:

High: 95-97%. Out of 100 2 are very close, 3 are wrong.

Medium: 94-98%. Out of 50 1 is lower/uppercase difference, 1 is just about correct (but there were much better words) and 1 is wrong.

Low: 68-72%. Out of 25, 4 are definitely wrong, 2 are close, 1 is half wrong, 1 is suspicious.

Though not perfect, the high and medium confidence files have matched well so these were imported to MLSN. They are flagged (with ja1h and ja1m respectively, in the comment field). Once an automatically added entry has been reviewed by a human expert it is either fixed, or deleted, or if it was correct then the code in the comment field is removed. (That is also the point the MLSN project takes ownership of the translation, see [1].)

The low confidence file scores around 70% which is fair but not good enough to justify importing them into MLSN.

The other thing that stands out from the word lists is that they are mostly obscure words. This is a direct result of not tackling polysemous words: generally in a language the polysemous words are also the most common words.

Bond [10] tackles the polysemy issue by cross-referencing with WordNet in other languages. 9,487 nouns matched via a WordNet in another language were matched, which is a similar size to the results presented here. In contrast to our results, 2,387 of those 9,487 were in the most common 3,300 nouns in the English language. However the downside of trying to automate the translation of polysemous nouns is relatively poor accuracy: 54.40% correct.

Conclusion

We have shown how a simple algorithm can give highly reliable results for the automated translation of the monosemous entries in a semantic network. The algorithm can be applied to any language provided suitable dictionaries are available.

References

[1]: Darren Cook. 2008. MLSN: A multi-lingual semantic network. In 14th Annual Meeting of the Association for Natural Language Processing. Tokyo.

[2]: <http://dcook.org/mlsn/>

[3]: Christine Fellbaum, editor, 1998. WordNet. An Electronic Lexical Database. MIT Press.

[4]: http://www.csse.monash.edu.au/~jwb/j_jmdict.html

[5]: <http://dict.tu-chemnitz.de/>

[6]: The work in this paper uses a 2005 download from the older version of CEDICT at <http://www.mandarintools.com/cedict.html>. There is a newer version at: <http://www.mdbg.net/cedictwiki/>

[7]: Harvesting Wikipedia Interwiki Links. <http://dcook.org/mlsn/about/>

[8]: http://en.wikipedia.org/wiki/Kenky%C5%ABsha's_New_Japanese-English_Dictionary

[9]: <http://www.alc.co.jp/> (ALC uses the Eijiro dictionary: <http://www.eijiro.jp/>)

[10]: 多言語 WordNet を利用した日本語 WordNet の作成 Francis Bond, 井佐原均, 神崎享子, 内元清貴 (2008). In 14th Annual Meeting of the Association for Natural Language Processing. Tokyo.

Appendices

ja1_high_sample:

15vina_del_mar0 ビニヤデルマル
04fourth_crusade0 第4回十字軍
06virtual_memory0 仮想記憶
06photomontage0 フォトモンタージュ
09fractal0 フラクタル

?????? 25dodecahedron0 正十二面体 ----> GG5 says this is "regular dodecahedron". Wikipedia says: "A dodecahedron is any polyhedron with twelve faces, but usually a regular dodecahedron is meant". However WordNet gloss is "any polyhedron having twelve plane faces".

18lucas0 ジョージ・ルーカス,ルーカス
06dead_axle0 死軸 --> Hard to track down, but see

<http://www.iwanami.co.jp/moreinfo/6021220/top.html>: "死軸 (デッド・アクスル) "

19gene_expression0 遺伝子発現
06sperm_bank0 精子バンク
04bohemianism0 ボヘミアニズム
24absentee_rate0 欠勤率
17ore_bed0 鉱層
13pound_cake0 パウンドケーキ
13penne0 ペンネ
15galveston0 ガルベストーン
15coahuila0 コアウイラ州
13cat_food0 キャットフード
18snoopy0 スヌーピー
18hunter0 猟師
07divisibility0 可分性
18clytemnestra0 クリュタイムネストラ
18agni0 アグニ
17lakeside0 湖岸
27brucine0 ブルシン
15balkan_peninsula0 バルカン半島
15uruguay0 ウルグアイ
26hepatitis_c0 c型肝炎
13gnocchi0 ニョッキ
15medan0 メダン
05affirmed0 アファームド --> Name of a horse!! But this is correct.
14camorra0 カモッラ
18aesop0 アイソーボス
06electric_clock0 電気時計
27tridymite0 鱗珪石
26car_sickness0 車酔い

***** 23kobo0 小保 --> WordNet is referring to Nigerian currency. 小保 comes from jmdict_places and jmdict_names.

25true_anomaly0 真近点角
21tax_liability0 納税義務
15auvergne0 オーベルニュ山地
08intervertebral_disc0 椎間板
26hepatitis0 肝炎
13brie0 ブリー

?????? 14string_orchestra0 弦楽合奏 --> GG5 says "a string ensemble". ALC says 弦楽合奏 is string concert, and that string orchestra is 弦楽合奏団. ja.wikipedia says 弦楽合奏 is a type of 合奏 and that 合奏 is a number of musicians performing together. In WordNet string orchestra refers to a type of musical organization: a group of instrumentalists. I therefore think this is wrong: 弦楽合奏団 is the organization.

15almaty0 アルマアタ,アルマトイ --> GG5 only gives the 2nd one, but WordNet gives both "Almaty, Alma-Ata" so I think this is correct

18mining_engineer0 鉱山技師
26obsessive-compulsive_disorder0 強迫性障害
06electric_locomotive0 電気機関車

08glomerulus0 糸球体
14metric_space0 距離空間
18hiawatha0 ハイアワサ
15dubuque0 ドゥビューク --> WordNet says it is iowa, ALC says it is illinois.
18boatbuilder0 舟大工 ---> Cannot find in any dictionary, but this article mentions it in the correct context: <http://ja.wikipedia.org/wiki/%E5%9D%AA%E5%B1%B1%E8%B1%8A>
**** 15omiya0 御宮 --> WordNet is referring to kanto placename, i.e. 大宮. 御宮 means shrine.
10lexeme0 語彙素
04civil_defense0 民間防衛
17epsilon_aurigae0 アル・マーズ
15north_cascades_national_park0 ノース・カスケード国立公園
15bryce_canyon_national_park0 ブライスキャニオン国立公園
08bull_neck0 猪首
04ecotourism0 エコツーリズム
18eris0 エリス --> GG5 lists Ellis, (Henry) Havelock. but also has the greek goddess which is the WordNet meaning here.
18robin_hood0 ロビン・フッド
11charybdis0 カリュブデイス
10film_noir0 フィルム・ノワール
**** 27town_gas0 都市ガス ---> WordNet refers to a type of coal gas. "town gas" is U.K. related. (Neither town gas nor coal gas are in en.wikipedia). ALC has coal gas as 石炭 {せきたん} ガス. ALC has 都市 {とし} ガス for town gas.
17great_barrier_reef0 グレート・バリア・リーフ
14intensive_care_unit0 集中治療室
18organization_man0 組織人間
06tokamak0 トカマク型 ---> ALC and GG5 say the Japanese is just "トカマク"? They don't have トカマク型 except as part of the word トカマク型臨界プラズマ試験装置. However Wikipedia says tokamak is トカマク型 and is definitely talking about the same thing. So I think this is correct.
18converso0 コンバルソ
18mennonite0 メノナイト
10dead_sea Scrolls0 死海文書
27sedimentary_rock0 堆積岩
05sir_barton0 サーバートン ---> Another horse! Correct.
15perihelion0 近日点
15mesa_verde_national_park0 メサ・ヴェルデ --> メサ・ヴェルデ国立公園 seems to be a synonym, but this still seems correct (e.g. according to ja.wiki)
19sleet0 霰
05pathogen0 病原体
15salina0 サライナ
05wagtail0 鶺鴒
06rotary_press0 輪転機
28revolutionary_calendar0 革命暦
13custard0 カスタード
15ciudad_juarez0 シウダードファレス,ファレス ---> ALC confirms that ファレス is the translation of Juarez meaning the place name (but does not explicitly say it is the same city in Mexico as シウダードファレス); GG5 confirms シウダードファレス. Wikipedia has no entry for ファレス, so unable to confirm it. Marked as correct for the moment.
15mindoro0 ミンドロ島
04war_of_the_spanish_succession0 スペイン継承戦争
06typesetting_machine0 植字機
15pisa0 ピサ
09mechanical_engineering0 機械工学
18music_critic0 音楽評論家
15greensboro0 グリーンズバラ
19bandwagon_effect0 バンドワゴン効果
28mesolithic_age0 亜旧石器時代,中石器時代 --> GG5 confirms 中石器時代 but has no entry for 亜旧石器時代. Ditto for ALC. ja.wiki says 亜旧石器時代 is Epipaleolithic: "中石器時代とも呼ばれ". en.wiki says "The term [Epipaleolithic] is sometimes confused with Mesolithic, and are sometimes used as synonyms."

Therefore marked as correct as WordNet does not differentiate.

08ileocecal_valve0 回盲弁
09counterculture0 カウンターカルチャー
06mercator_projection0 メルカトル図法
06workroom0 仕事部屋
09orientalism0 東洋学
14building_society0 住宅金融組合

ja1_medium_sample

10appellation0 称呼
18latrobe0 ラトローブ ---> ALC tells me this is fine for the person's name. It doesn't specifically mention Benjamin Henry Latrobe, but is still correct.
05madagascar_cat0 ワオキツネザル
06canopic_jar0 カノプス壺
20radish3 ダイコン ---> Yes, 20radish3 is "radish, daikon, Japanese radish, Raphanus sativus longipinnatus"

20betulaceae0 カバノキ科
15sierra_leone0 シエラレオネ
***** 14choir_school0 スコラ・カントルム ---> schola cantorum in English can apparently refer to any choir school. But in Japanese it refers to "La Schola Cantorum is a private music school in Paris."

Correct Japanese for 14choir_school0 is 聖歌隊付属学校

18thea0 テイアー
06brooch0 ブローチ,胸飾り
18positivist0 実証主義者
08tendon0 腱
18haydn0 フランツ・ヨーゼフ・ハイドン
18vancouver0 ジョージ・バンクーバー
27vitamin_d0 ビタミン d

?????? 06ace_inhibitor0 ace 阻害薬 ---> ace should be in uppercase in the Japanese, but lowercasing words sourced from wikipedia is a design decision. Otherwise it is correct.

?????? 18retailer0 購入先 ---> Not in GG5 or ALC or ja.wiki. On the other hand given this word native speakers understand it to mean a shop (or a firm selling something). Better Japanese synonyms: 小売業者[こうりぎょうしゃ], 小売店[しょうばいてん], 小売業[こうりぎょう]. It seems like 購入先 may actually belong to something else in WordNet, but I'm not clear what...

20lentibulariaceae0 タヌキモ科
17coast_range0 コースト山脈 --> Correct, both are referring to western U.S.
18rameses0 ラムセス
05brine_shrimp0 アルテミア
04observation0 観望 ---> Observation is polysemous but 観望 does seem to be this "the act of observing; taking a patient look" meaning, especially as stargazing (天体観望) is a hyponym.

18cherubini0 ルイジ・ケルビーニ
04aerobics0 有酸素運動
27hydrocortisone0 コルチゾール
20acanthaceae0 キツネノマゴ科
26leishmaniasis0 リーシュマニア症
05little_auk0 ヒメウミスズメ
18nero0 ネロ ---> This referring to the Roman Emperor, and is correct (both nero and ネロ have other meanings though)

05procellariiformes0 ミズナギドリ目
23mark0 ドイツマルク
18eindhoven0 ウィレム・アイントホーフェン
20oregon_cedar0 ローソンヒノキ
23ugandan_shilling0 ウガンダ・シリング
20florist's_gloxinia0 グロキシニア
06steamroller0 ロードローラー
04conditional_reflex0 条件反射
18villa-lobos0 エイトル・ヴィラ=ロボス
16inner_light0 内なる光

23parsec0 パーセク
27propenonitrile0 アクリロニトリル
18goodman0ベニー・グッドマン,グッドマン
27neodymium0 ネオジウム
18henry0 ジョセフ・ヘンリー
20japanese_quince0 木瓜
04social_welfare0 公的扶助,社会福祉 ----> 公的扶助 is public assistance according to both GG5 and ALC. Ah, that meaning is covered by the WordNet entry. So this is correct.
28new_year's_eve0 12月31日
08nasal0 鼻骨
14committee_for_state_security0 ソ連国家保安委員会
20typha0 ガマ

ja1_low_sample

07hatefulness0 憎さ
10bracket0 鉤括弧 ---> Note: 「」 in Japanese are also called this
15east_saint_louis0 イーストセントルイス
**** 05botulinus0 ボツリヌス中毒,ボツリヌス菌 ---> ボツリヌス中毒 is botulism. ツリヌス菌 is correct.
23boltzmann's_constant0 ボルツマン定数
26dankness0 厥冷 ---> 厥冷 is not in dictionaries. But this reference confirms it can be dankness, so I'm marking it as correct: <http://www.websters-online-dictionary.org/definition/clamminess>. But 湿っぽさ is better?
21certificate_of_deposit0 譲渡性預金 --> Also called 譲渡可能定期預金証書
17constance0 ボーデン湖
***** 15thorshavn0 トルスハウン ---> ALC gives this as a place name (nothing specific), but wikipedia says capital of Faroe islands is トースハウン. So I think this one is wrong.
***** 13frosting0 霜降り ---> WordNet is talking about icing, which is "アイシング, 糖衣[トウイ]". 霜降り appears to be a marbling effect and/or 06pepper-and-salt0.
18ironworker0 鉄工 --> GG5 agrees, ALC says 鉄工具, but I think this is correct.
06buffer1 研磨機 ---> WordNet says "a power tool used to buff surfaces". ALC says "研磨機" is abrasive machine // grinder // grinding machine // mill // polishing machine // sanding machine. GG5 says it is "a grinding machine; a grinder; an abrader; [レンズなどの] a polishing machine; a polisher". Wikipedia does not comment. <http://www.fritsch.co.jp/polishingmachine.html> shows polishing machines for glass. I've decided to mark this as correct, assuming 研磨機 is polysemous for both 06buffer1 and 06grinder0.
***** 18wing_commander0 少佐 ----> WordNet says "(RAF rank) one who is next below a group captain". ALC says wing commander is a British term for 空軍中佐, and that 少佐 is: gold oak leaf // lieutenant commander 《海軍》 [【略】 Lt. Com.] // major 《軍事》 // squadron leader 《英空軍》. Group captain is 大佐.
06open-air_market0 青空市場
08islands_of_langerhans0 ランゲルハンス島
10chortle0 含み笑い
***** 20salal0 シャロン ----> "salal, shallon, Gaultheria shallon" is a small evergreen shrub. <http://blog.livedoor.jp/ht73101/> says it is "レモンリーフ" or "ゴータリア・シャロン".
????? 06microbrewery0 地ビール ---> The interwiki link matches, and they are both talking about the same thing. But is microbrewery actually "地ビールの醸造所"?
26amblyopia0 弱視
18in-law0 姻族
18gilman0 ギルマン ---> "Charlotte Anna Perkins Gilman". But ギルマン is valid for Gilman so marked as correct.
08spicule0 骨片
***** 11vroom0 ブルーム ---> This is the sound of an engine. The closest ブルーム seems to be the "Vroom" movie. The correct Japanese is "ブルーン、ブンブン".
***** 10intercourse0 社交 ---> Close, but I think 社交 is being sociable, whereas "social intercourse" is talking about the actual communication. Correct is maybe: 交際, 交流
15gadsden0 ガズデン