WordNet – лексична база даних англійської мови

WordNet, це семантично орієнтований словник англійської мови, подібний до традиційних тезаурусів але з більш багатою структурою. У *WordNet* слова групуються у набори синонімів – синсети, кожен із своїм визначенням і зв'язками з іншими синсетами. *WordNet 3.0* розповсюджується разом з NLTK і містить 155287 слів та 117659 синсетів. Хоча *WordNet* розроблявся для психолінгвістики - цей словник широко використовується в NLP та в задачах інформаційного пошуку. Розробки для інших мов проводяться на основі документації, яка наведена у http://www.globalwordnet.org/.

Словник можна запустити як окремий об'єкт за адресою http://wordnetweb.princeton.edu/perl/webwn. Він відкривається у вікні браузера, має зручний для розуміння і використання інтерфейс.

Значення і синоніми

Розглянемо наступне речення:

```
(1) Benz is credited with the invention of motorcar.
```

Якщо замінити слово `motorcar` на `automobile` зміст речення не зміниться.

```
(2) Benz is credited with the invention of automobile .
```

Можна вважати, що оскільки заміна слів не вплинула на зміст речень то ці слова синоніми. Для одержання значення слова потрібно вибрати до якої частини мови воно належить. `WordNet` містить чотири словники (іменники, дієслова, прикметники, прислівники). Знайдемо слово `motorcar` у словнику іменників:

Слово `motorcar` має одне можливе значення і воно ідентифікується як перший сенс іменника `car`. `car` називають синсетом – множиною синонімічних слів.

Слова в синсет об'єднані за спільним значенням, яке є однакове для всіх слів. В синсеті вказується текстовий опис цього значення та приклад вживання слів з синсету.

Слова `automobile` та `motorcar` є однозначні і входять тільки в один синсет. Слово `car` багатозначне і входить в п'ять синсетів.

Ієрархія в WordNet

Синсети відповідають абстрактним поняттям, які можуть мати або не мати відповідних слів. Ці поняття зв'язуються разом в ієрархії. Деякі поняття Entity, State, Event – є загальними і їх називають унікальними початковими поняттями. Інші є більше специфічними. Частина ієрархії понять наведена на рис.4. Лінії між вузлами вказують на зв'язки (гіперонім\гіпонім), пунктирна лінія вказує, що *artefact* не є безпосереднім гіперонімом *motorcar*.

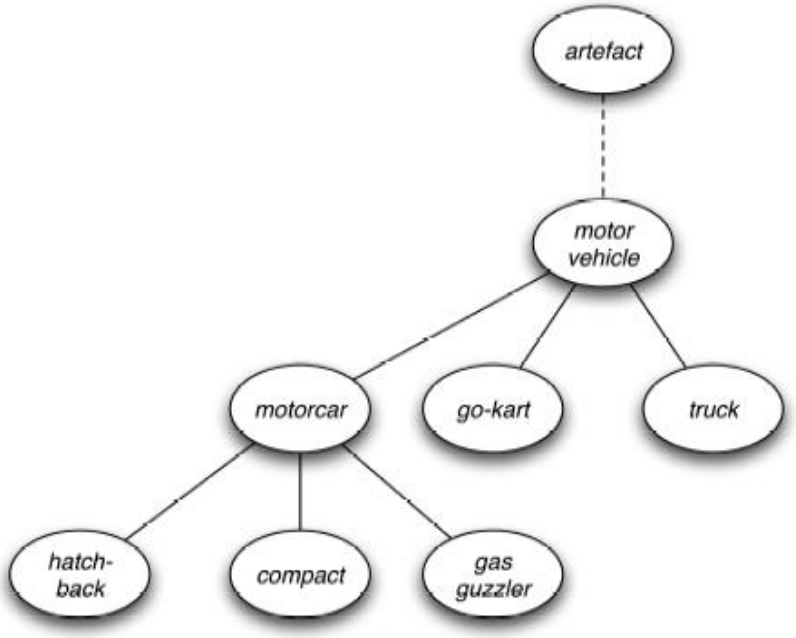

Рис.4. Фрагмент ієрархії понять

WordNet дозволяє легко переміщатися між поняттями. Наприклад для поняття *motorcar* ми можемо переглянути поняття, які є більш специфічними (гіпонім):

Аналогічно можна піднятися по ієрархії і переглянути більш широкі поняття ніж motorcar (гіперніми). Деякі слова мають декілька шляхів вверх, ці слова можуть класифікуватися більш ніж одним способом.

Лексичні зв'язки в WordNet.

Таблиця 1 містить список найбільш важливих типів зв'язків, які реалізовані у WordNet. Таблиця 2 містить повний список зв'язків іменників.

Таблиця 1

Hypernym	Узагальнення	Тварини є гіпернімом собаки
Hyponym	Деталізація	Собака є гіпонімом тварин
Meronym	Частина від	Двері є меронімом будинку
Holonym	Містить складові	Будинок є холонімом дверей
Synonym	Подібне значення	Машина є синонімом автомобіля
Antonym	Протилежне значення	Подобається є антонімом не подобається
Entailment	Необхідна дія	Крок є ентейлментом ходи

Таблиця 2

Relation	**Also Called**	**Definition**	**Example**
Hypernym	Superordinate	From concepts to	Breakfast →meal

		superordinates	
Hyponym	Subordinate	From concepts to subtypes	Meal → Lunch
Instance Hypernym	Instance	From instances to their concepts	Austen → Author
Instance Hyponym	Has-Instance	From concepts to concepts instances	Composer → Bach
Member Meronym	Has-Member	From groups to their members	Faculty → Professor
Member Holonym	Member-Of	From members to their groups	Copilot → Crew
Part Meronym	Has-Part	From wholes to parts	Table → Leg
Part Holonym	Part-Of	From parts to wholes	Course → Meal
Substance Meronym		From substances to their subparts	Water → Oxyden
Substance Holonym		From parts of substances to wholes	Gin → Martini
Antonym		Semantic opposition between lemmas	Leader ←→ Follower
Derivationally Related Form		Lemmas w/same morphological root	Destruction ←→ Destroy

Гіперніми та гіпоніми називають лексичними зв'язками тому що вони пов'язують один синсет з іншим. Ці два зв'язки вказують на рух вверх-вниз в ієрархії «is-a». Інший можливий шлях в ієрархії WordNet це від предмету до його складових (меронім), або до поняття яке містить предмет в собі (голоніми). Наприклад, частини дерева – стовбур, крона та ін. `part_meronyms()` . Речовина з якого дерево зроблено включає `heartwood` та `sapwood`; - `substance_meronyms()`. Багато дерев утворюють ліс - `member_holonyms()`

Оцінка подібності в WordNet

Синсети зв'язані між собою складною мережею лексичних зв'язків. Для певного синсету можна переглянути зв'язки у WordNet і знайти синсети, які з ним зв'язані за змістом. Інформація про семантичні взаємозв'язки між словами цінна при класифікації текстів.

Кожен синсет має один або більше шляхів за яким він зводиться до ключового поняття. Два синсети можуть мати спільне ключове поняття і чим нижче за ієрархією це ключове поняття тим ближчі між собою ці два синсети.

Семантична подібність двох понять пов'язана з довжиною шляху між цими поняттями в *WordNet* . Пакет *wordnet* містить багато засобів для здійснення таких вимірювань (Leacock-Chodorow, Wu-Palmer, Resnik, Jiang-Conrath, Lin). Наприклад *path_similarity* (присвоює значення від 0 до 1) базується на найкоротшому шляху, який поєднує поняття за ієрархією гіперонімів (-1 означає що шлях (спільний гіперонім) не знайдено).