Caroline BARRIÈRE, School of Information Technology and Engineering, University of Ottawa,
Dan FASS, School of Computing Science, Simon Fraser University, Burnaby

Dictionary validation through a clustering technique

Abstract

Barrière and Popowich (1996) have developed a technique, implemented by Barrière (1997), that takes as a "trigger" a word sense defined in a dictionary and finds its **concept cluster** – a group of word senses related to the trigger that shows its larger semantic context in the dictionary. The technique can be extended to take a concept cluster as a "trigger" and find an **extended concept cluster** by combining the clusters found for the word senses in the trigger cluster. In this paper, we present two applications of the concept clustering technique to validating dictionary definitions. The first application is a qualitative tool for seeing connections and noting inconsistencies among dictionary definitions; the second application is two quantitative measures of the coherence of dictionary definitions and interdefinitions. The first measure can help in writing tight-knit interdefinitions of related word senses; the second can help ensure that individual word sense definitions are of high quality.

Keywords: Concept clustering, conceptual graphs, evaluation of dictionary definitions

1. Introduction

Barrière and Popowich (1996) have developed a dictionary analysis technique, implemented in the ARC-Concept (Acquisition-Representation-Clustering of Concepts) system (Barrière 1997), that takes the dictionary definition of a "trigger" word sense and builds its **concept cluster**. A concept cluster is a group of concepts that comprise a micro-world of objects, actions, and participants related to the trigger. The technique can also use a concept cluster as its "trigger" and produce an **extended concept cluster** based on the word senses that were part of the trigger cluster.[1]

The clustering technique helps show the significant semantic content in dictionary definitions. In this paper, we suggest that it can be embodied in two applications that help validate dictionary definitions. The first application is a tool for better understanding the semantics of the dictionary-definition writing process and for critiquing the semantics of particular definitions. We think that showing lexicographers a concept cluster and its extended concept cluster throws light on the definitions of the word senses involved and allows lexicographers to see connections and note inconsistencies among their definitions.

The second application is two quantitative measures of the coherence of dictionary definitions and interdefinitions, also based on the concept clustering technique. The first measure quantifies the interrelationship between dictionary definitions of "related" word senses that are part of the same cluster. Lexicographers can use this measure to ensure that concept clusters are tight-knit. The second measure quantifies the quality of a dictionary definition in relation to its extended cluster. Lexicographers can use this measure to validate individual dictionary definitions in relation to the definitions of other word senses in an extended cluster.

The paper is organized as follows. In section 2 and section 3, the basic clustering technique and the extended clustering technique are described for generating respectively concept clusters and extended concept clusters. Results from the ARC-Concept system are presented and discussed. In section 4, two applications of the clustering technique to dictionary validation are presented: 1) the qualitative tool for understanding and critiquing dictionary definitions and 2) two measures of the coherence of dictionary definitions and inter-definitions.

2. Basic clustering technique

The clustering technique, explained in detail in (Barrière 1997), has been implemented in the ARC-Concept system. The system uses an electronic version of *The American Heritage First Dictionary* (AHFD)[2], which is for children aged 6-8 and contains 1800 dictionary definitions. ARC-Concept automatically translates the definitions into Conceptual Graphs (CGs) (Sowa 1984), hence aiming at a CG representation of each definition with nodes being concepts (identified from the words used in the definition) and links are relations (role relationships between those concepts). Figure 2.1 shows the definition of the word ***airplane*** from the AHFD and its CG representation.

Word: ***airplane***
Definition: An airplane is a machine with wings that flies in the air.
Airplanes carry people from one place to another.

```
[machine:a]-                           [carry]-
  {                                       {
   (with)→[wing:plural];                    (object)→[person:plural];
   (agent)←[fly]→(in)→[air:the];            (from)→[place:one]→(to)→[another:ref];
   (is-a)←[airplane:an];                    (agent)→[airplane:plural];
  }.                                      }.
```

Figure 2.1 - Example AHFD definition and its Conceptual Graph representation

A **concept cluster** starts with a single element: a trigger concept (such as ***airplane*** in Figure 2.1). The CG representation of the concept cluster, which is called the **Concept Clustering Knowledge Graph** (CCKG), starts with a single element as well: the CG representation of the trigger concept. A **trigger phase** and an **expansion phase** produce a concept cluster containing more concepts and a larger CCKG to represent all the combined definitions of the concepts in the concept cluster.

For example, if ***airplane*** is the trigger, its CG definition (shown in Figure 2.1) becomes the starting CCKG which might be joined to the CG definitions of word senses such as ***wing***, ***air*** and ***machine*** (all words with a single sense) or to senses of words such as ***fly*** (which has 2 senses) to form a concept cluster ***{wing, air, machine, fly_2}*** which will have a large CCKG for its representation.

A word sense or word must fulfill a **Frequency Threshold** (FT) criteria to become a candidate for joining the CCKG. The frequency of occurrence of all words in the dictionary is calculated. Because it is not possible to distinguish between the possible senses of all words when computing those frequencies[3], then we work at the word level (instead of the concept level) and refer to the chosen candidates as **Semantically Significant Words**. A SSW has fulfilled the FT criteria by occurring less than a fixed number of times in the dictionary.

Table 2.1 shows some statistics from the AHFD on frequency thresholds. For example, a FT set at 27 means that only words occurring less than 27 times in the dictionary will be considered as SSWs. This results in eliminating 13% of all the possible words in the dictionary, which account for 80% of all occurrences of words in the dictionary. From Table 2.1, we can also establish that 60% of all words in the dictionary occur four times or less.

Table 2.1 - Frequency Thresholds

FT	% of occurrences discarded	% words discarded
5076	10	< 1
2904	20	< 1
1095	30	< 1
623	40	< 1
264	50	1
126	60	2
62	70	6
42	75	8
35	77	10
27	80	13
18	85	19
11	90	29
5	95	48
4	97	60
2	99	80
1	100	100

Back to the example from Figure 2.1, the FT would be used to select as SSWs, such words as ***wing*** and ***fly*** from the definition of ***airplane***, but to exclude those such as ***person*** or ***place***, which are also part of the definition. The latter are so frequent that they do not contribute significant semantic content.

The graph representation of the candidate SSWs must then fulfill a second criteria: a **Graph Matching Threshold** (GMT). The GMT determines whether a SSW is suitable for joining the concept cluster. The GMT consists of two values: (1) the number of SSWs shared by the CCKG and the CG representation of the candidate SSW, (2) the number of relations (i.e., role relationships) shared by the CCKG and the CG representation of the candidate SSW; these relations have to link shared concepts to lead to common subgraphs.

Setting the FT and the values in the GMT are a matter of experimentation.

The trigger phase of the clustering process builds an initial concept cluster around one source: the trigger word sense. The trigger phase consists of trigger forward and trigger backward steps. At any point in the trigger phase, when the CG representation of a word sense of a candidate SSW is joined to the CCKG, that word sense becomes part of the concept cluster.

Trigger forward:

1) Find the SSWs in the CCKG (the initial CCKG is the CG representation of the trigger's definition) and
2) attempt to join their respective CGs to the CCKG based on the GMT.

Trigger backward:

1) Find all SSWs in the dictionary that use the trigger in their definition and
2) attempt to join their respective CGs to the CCKG based on the GMT.

The GMT is set low for the trigger phase: GMT(1,0), i.e., a single SSW in common and no relations, which means that the graph to be joined must have one SSW in common with the CCKG.

The expansion phase enlarges the concept cluster by investigating all the SSWs now part of the CCKG. The aim of the expansion phase is to create a more interconnected graph rather than expanding from a particular word sense. For this reason, a GMT higher than the one in the trigger phase is used. The expansion phase consists of expansion forward and backward steps. Again, at any point in the expansion phase, when the CG representation of a word sense of a candidate SSW is joined to the CCKG, that word sense becomes part of the concept cluster.

Expansion forward:

1) Find the SSWs in the CCKG and
2) attempt to join their respective CGs to the initial CCKG based on the GMT.

Expansion backward:

1) Find all SSWs in the dictionary that use any SSW from the concept cluster in their definitions and
2) attempt to join their respective CGs to the CCKG based on the GMT.

Repeat:

Continue forward and backward expansion until no further changes are made (no new SSWs are becoming part of the cluster).

The words considered as SSWs are verified to see whether they have multiple senses or not. If a word has multiple senses, the clustering process attempts to join the CG representations of all its senses. To discourage incorrect word senses from being included in the concept cluster, the GMT is raised. This also makes it harder for the correct word sense to become part of the concept cluster, but we make the assumption that it is better to not include a correct word sense than include incorrect ones.

A resulting concept cluster will consist entirely of concepts (SSWs with a single word sense, and chosen senses of SSWs with multiple senses), and the graph representation of all those concepts put together forms the CCKG which is the graph representation of the concept cluster.

Table 2.2 shows some examples of concept clusters that result from using the clustering technique around a different trigger (word senses). Column 1 lists the trigger, column 2 shows the Frequency Threshold (FT), column 3 shows the Graph Matching Threshold (GMT), and column 4 shows the resulting concept cluster.

The results show that the number of word senses related to the trigger varies with the values set for the FT and GMT in columns 2 and 3 of Table 2.2. The effect of varying the FT (column 2) can be significant, as seen in the variation in the size of the clusters for ***needle_1*** and ***sew*** when the FT is 42 versus other values. However, it also sometimes has little effect, as seen in the two results for ***wash***.

Table 2.2 - Multiple concept clusters from different trigger word

Trigger concept	Frequency Threshold for SSW	GMT for expansion phase	Concept Cluster
needle_1	42	GMT(2,1)	10 words: {needle_1, sew, cloth, thread, wool, handkerchief, pin, ribbon, string, rainbow}
	18	GMT(3,0)	3 words: {needle_1, sew, thread}
	12	GMT(1,1)	2 words: {needle_1, thread}
sew	42	GMT(2,1)	16 words: {sew, cloth, needle_1, needle_2, thread, button, patch_1, pin, pocket, wool, ribbon, rug, string, nest, prize, rainbow}
	18	GMT(1,1)	6 words: {sew, needle_1, needle_2, thread, button, pocket}
	18	GMT(3,0)	6 words: {sew, needle_1, needle_2, thread, button, pocket}
soap	42	GMT(1,1)	12 words: {soap, help, dirt, mix, bath, bubble, suds, wash, clean_2, boil, anchor, steam}
	18	GMT(1,1)	9 words: {soap, dirt, mix, bath, bubble, suds, wash, boil, steam}
	18	GMT(3,0)	8 words: {soap, dirt, mix, bath, bubble, suds, wash, boil}
	8	GMT(1,1)	7 words: {soap, bath, bubble, suds, wash, boil, steam}
wash	42	GMT(1,)	9 words: {wash, soap, bath, bathroom, suds, bubble, boil, clean_2, steam}
	16	GMT(1,1)	8 words: {wash, soap, bath, bathroom, suds, bubble, boil, steam}

The clusters in Table 2.2 should not be confused with thesaurus entries. A set of words in a thesaurus entry have senses that all share a particular aspect of their meaning. For example, ***farmer*** is related to the set ***{husbandman, horticulturist, gardener, florist, agricultor, agriculturist, yeoman, cultivator}*** because all are subclasses of "people working with the soil". A formal concept analysis (Wille 1997) of this set of words would show they have one or more features in common. On the other hand, a set of word senses forming a concept cluster, as defined in this paper, do not share a set of features, but instead share a common micro-world or context.

3. Extended clustering technique

The clustering technique can also be used with a cluster as the start-point, which we call a **trigger cluster**. The result is an **extended concept cluster** (or "cluster of clusters"). Two examples (Figure 3.1 and 3.2) illustrate this use of the clustering technique. In Figure 3.1, the trigger concept ***post-office*** is used to generate the concept cluster ***{post-office, mail_2, address, package, mail_1, stamp}*** and then each concept, member of the cluster, is itself used as a trigger to generate a concept cluster. The extended concept cluster is the result of joining all concept clusters into one larger cluster.

trigger concept: **post-office**
trigger cluster: **{address, mail_1, mail_2, package, post-office, stamp}**

Individual concept clusters:

1. address: {address, mail_1, mail_2, package, post-office, stamp}
2. mail_1: {address, mail_1, mail_2, package, post-office, stamp}
3. mail_2: {mail_2, mail_1, package, stamp}
4. package: {address, bring, mail_1, mail_2, package, post-office, stamp}
5. stamp: {address, letter_1, mail_1, mail_2, package, post-office, stamp}

Extended cluster: {mail_1, mail_2, package, stamp, address, post-office, bring, letter_1}

Trigger cluster (6 concepts) and extended cluster produced (8 concepts):

Description of following three columns:
[1] concepts present in the extended cluster,
[2] number of times a concept occurs (calculated from the 6 individual clusters),
[3] % of times a concept occurs (calculated from the 6 individual clusters).

[1]	[2]	[3]
mail_1, mail_2, package, stamp	(6)	(100% = 6/6)
address, post-office	(5)	(83.3% = 5/6)
bring, letter_1	(1)	(16.7% = 1/6)

Figure 3.1 - Extended concept cluster derived from trigger ***post-office***

In Figure 3.2, we show the extended cluster resulting from using ***farmer*** as the trigger. The Frequency Threshold for SSWs was lowered to reduce the search space, but the Graph Matching Threshold was lowered as well (to GMT(1,0) trigger, GMT(2,0) expansion) to allow more matches.

trigger concept: ***farmer***
trigger cluster: ***{cowboy, cowgirl, farm, farmer, goat, raise_2, tractor}***

Individual concept clusters:

1. cowboy: {cattle, cowboy, cowgirl, farm, farmer}
2. cowgirl: {cattle, cowboy, cowgirl, farm, farmer}
3. farm: {barn, cattle, cow, cowboy, cowgirl, farm, farmer, goat, pig, raise_2, stable}
4. goat: {chin, cotton, farm, goat, hair, lion, lioness, raise_2, stable, sweater, tugboat, wizard, wool, yarn}
5. raise_2: {cowboy, cowgirl, farm, farmer, raise_2}
6. tractor: {engine, farmer, prepare, tractor, tugboat}

Extended cluster: {farm, farmer, cowboy, cowgirl, raise_2, cattle, goat, stable, tractor, tugboat, barn, chin, cotton, cow, engine, hair, lion, lioness, pig, prepare, sweater, wizard, wool, yarn}

Trigger cluster (7 concepts) and extended cluster produced (24 concepts):

Description of following three columns:
[1] concepts present in the extended cluster,
[2] number of times a concept occurs (calculated on the 7 individual clusters),
[3] % of times a concept occurs (calculated on the 7 individual clusters).

[1]	[2]	[3]
farm, farmer	(6)	(85.7% = 6/7)
cowboy, cowgirl	(5)	(71.4% = 5/7)
raise_2	(4)	(57.1% = 4/7)
cattle, goat	(3)	(42.8% = 3/7)
stable, tractor, tugboat	(2)	(28.5% = 2/7)
(all others)	(1)	(14.2% = 1/7)

Figure 3.2 - Extended concept cluster derived from trigger *farmer*

4. Applications to dictionary validation

The basic clustering technique and extended clustering technique can be viewed as showing how the significant semantic content in dictionary definitions is provided. That content can be understood in terms of concept clusters and CCKGs that contain Semantically Significant Words selected by the use of Frequency Thresholds and Graph Matching Thresholds. These ideas have applications to the validation of dictionary definitions and interdefinitions (i.e., definitions of semantically related word senses).

Two applications of these ideas are presented below. The first application is a qualitative tool for seeing connections and noting inconsistencies among dictionary definitions. The second is two quantitative measures of the coherence of dictionary definitions and interdefinitions.

4.1. Visualization of dictionary definitions and interdefinitions

Both the basic and extended clustering techniques can help lexicographers "visualize" or "understand" better the semantic content of dictionary definitions. The basic clustering technique can help lexicographers evaluate, when starting from one word sense, which others are used to build a concept cluster.

The basic clustering technique, if enhanced with a visualization tool, would allow the lexicographers to qualitatively inspect individual definitions for potential inconsistencies. For example, Table 2.2 shows that ***soap*** reaches ***{bath, bubble, suds, wash, boil, steam}***. A lexicographer might ask why ***soap*** reaches ***boil*** but not ***shower*** or ***laundry*** which seem more semantically related to ***soap***.

The extended clustering technique produces sets of clusters from an individual trigger cluster. The sets of clusters give more information about which word senses seem to be interdependent and co-occur. It provides a way to specify how those senses might be "defined in a similar manner".

The extended clustering technique, if similarly enhanced with a visualization tool, would allow the lexicographers to check whether related word senses were similarly defined, i.e., that they use in their respective definitions a common group of word senses. Using such a technique, a lexicographer could also find missing links and inconsistencies between word senses. The extended concept cluster of ***farmer*** shown in Figure 3.2, for example, raises questions such as why ***cattle*** is often reached along with ***farm*** but not ***chicken*** and why ***cowboy*** and ***cowgirl*** are always reachable along with ***farm*** and ***farmer***.

4.2. Dictionary coherence

We have developed two measures of "coherence" that apply to dictionaries.

Measure 1: Tight-knittedness of extended cluster

This measure – simply the number of concepts in the extended cluster relative to the trigger cluster – helps quantify the quality of extended clusters produced from trigger clusters and could be used by lexicographers to improve the quality of interdefinition of a group of related concepts such as ***post-office***, ***package***, and ***address***.

Intuitively, the extended cluster for ***post-office*** shown in Figure 3.1 is more "coherent" than the extended cluster for ***farmer*** shown in Figure 3.2 and this measure bears out the intuition: there are 8 concepts in the extended cluster for ***post-office*** but a much higher 24 concepts for ***farmer*** (the sizes of their respective trigger clusters are very similar at 6 and 7, respectively).

The higher numbers for ***farmer*** have a number of interpretations. Among them are the following 3 interpretations that the domain for ***farmer*** (1) expresses a more complex micro-world than that for ***post-office***, (2) has weaker definitions containing (more) unnecessary or inconsistently used concepts, and (3) is affected by other, partially-overlapping micro-worlds. Interpretation (2) is seen in Figure 3.2 in the inconsistent use of ***cattle***, ***goat*** and ***sheep***. Interpretation (3) can help explain the large size of the extended cluster around ***farmer*** with ***goat*** being particularly significant: it alone introduces 9 concepts into the extended cluster.

Measure 2: Quality of individual definitions in relation to extended cluster

This measure quantifies how well each definition conforms to its extended cluster. Lexicographer could use the measure to isolate definitions that don't conform well and write better versions. Let us apply the measure to the extended cluster around ***post-office*** presented in Figure 3.1. The cluster for ***mail_2*** conforms 100% with the extended cluster. This result is computed as follows: (4 x 1) = 4 out of a possible 4 = 100%, i.e., the 4 concepts in the definition of mail_2 are in all 6 subclusters in the extended cluster. The clusters for ***address***, ***mail_1***, and ***post-office*** all conform 94.4% with the extended cluster. This is computed as: (4 x 1) + (2 x 0.833) = 5.666 out of a possible 6 = 94.4%, i.e., 4 concepts occur in all 4 subclusters and 2 concepts occur in 5 of the 6 subclusters. By contrast, the clusters for ***package*** and ***stamp*** conform 83.3%. This is computed as: (4 x 1) + (2 x 0.833) + (1 x 0.167) = 5.833 out of a possible 7 = 83.3%.

The trigger concepts in Figure 3.2 can also be so rated, with the following results: ***raise_2*** (74.2%), ***cowboy*** (71.4%), ***cowgirl*** (63.2%), ***farm*** (55.7%), ***tractor*** (31.3%), and ***goat*** (26.3%).

The numbers quantify the intuition that the definitions in the ***post-office*** micro-world are more alike than those for the ***farmer*** micro-world. They also quantify the intuition that certain individual definitions are better than others with respect to a common extended cluster, e.g., the definition for ***raise_2*** is better than that for ***goat*** with respect to the extended cluster for ***farmer*** in Figure 3.2.

5. Conclusion

We have presented a concept clustering technique that may help lexicographers validate dictionary definitions and interdefinitions by helping them decide whether they are of an acceptable standard or not. The clustering technique, implemented within ARC-Concept, embedded in a visualization tool (yet to be implemented) may allow lexicographers to better understand the significant semantic content in dictionary definitions, see hitherto unnoticed links among definitions, and also note discrepancies between them. This could help lexicographers write better dictionary definitions.

The clustering technique also provides the basis for two coherence measures. The first measure is of the tight-knittedness of the interdefinition of related word senses (in an extended concept cluster). The second measure is of the quality of the dictionary definition of an individual word sense in relation to others with which it shares a common context (in an extended concept cluster). The first measure could be used to help write tight-knit interdefinitions of related word senses; the second could be used to help improve the quality of individual definitions.

6. Notes

[1] The term **word sense**, **word** and **concept** are used in this paper in the following ways. Some words defined in our application dictionary have multiple word senses. When the system detects such words, it lists their separates senses, e.g., *mail_1* and *mail_2* which are considered as two word senses. Other words in our application dictionary have only a single sense. They are written using the word form but are considered as word senses as well, e.g. *toe* is a word sense or a word with a single sense. The term concept is used interchangeably with word sense. Concept clusters, as shown in Table 2.2, Figure 3.1 and Figure 3.2 in the paper are formed entirely of word senses (concepts). In some cases, the system handles words without knowing (at that point) whether a word has one or multiple senses or without being able (at that point) to distinguish between the multiple senses of a word known to have more than one sense. Those words appear in Conceptual Graph representations of definitions and in Concept Clustering Knowledge Graphs and can become candidate "Semantically Significant Words" (see section 2).

[2] Copyright © 1994 by Houghton Mifflin Company. Definition reproduced by permission from *The American Heritage First Dictionary.*

[3] A possibility is to do a manual sense tagging of all words used in the dictionary, but this totally defeats the idea of automatic processing promoted in this research.

7. References

Barrière, Caroline (1997) *From a Children's First Dictionary to a Lexical Knowledge Base of Conceptual Graphs*, PhD thesis, Simon Fraser University.

Barrière, Caroline and Fred Popowich (1996) "Concept clustering and knowledge integration from a children's dictionary", In *Proc. of the 16th COLING*, Copenhagen, Danemark.

Sowa, J. (1984) "Conceptual Structures in Mind and Machines", Addison-Wesley.

Wille, R. (1997) "Conceptual graphs and formal concept analysis", In D.Lukose, H.Delugach, M.Keeler, L.Searle, and J.Sowa, editors, *Conceptual Structures: Fulfilling Peirce's Dream*, pages 290–303. Springer.