Information Engineering Express
International Institute of Applied Informatics
2015, Vol.1, No.4, 21 – 31

Vector Similarity of Related Words and Synonyms in the Japanese WordNet

Takuya Hirao[*], Takahiko Suzuki[†],
Nao Wariishi[‡], Sachio Hirokawa[§]

Abstract

Word2vec is a tool that produces vector representations of words from a large amount of text data. In this study, we show that only a part of the vector space produced by word2vec is sufficient to represent the collective sense of a set of related words in Japanese WordNet. Furthermore, we show that there is a subspace of the vector space that does not relate to the collective sense of related words and synonyms. We construct a compact decision tree by using the vectors to distinguish whether a given word belongs to the set of related words.

Keywords: Decision tree, Thesaurus, Word2vec, WordNet

1 Introduction

Approximately 5% of the entries in the Japanese WordNet [1, 2] contain errors, many of which are classified as “wrongly placed synonyms in a Synset.” Hirao et al. [3] proposed methods for detecting such errors. In previous studies, we attempted to detect errors based on a method that used the cosine similarity (CS) between vectors generated by word2vec [4]. However, the result was far from satisfactory. Based on our experience and other work on Word2vec [5, 6], we posed the following two hypotheses:

- Word2vec vectors do indicate the sense of synonyms in a Synset.

- There is “noise” in the vector spaces that diminishes the usefulness of CS values.

In this study, we attempt to verify these hypotheses. We will analyze a set of related words in hypernym–hyponym-related Synsets instead of synonyms in a Synset because the number of synonyms in a Synset is typically too small for statistical evaluation. The Synsets are selected from the Japanese WordNet in a manner that satisfies some conditions to be stated later. We assume that the resulting set of related words has a “collective sense.”

[*] Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan
[†] Research Institute for Information Technology, Kyushu University, Japan
[‡] Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan
[§] Research Institute for Information Technology, Kyushu University, Japan

Compact decision trees are created for distinguishing related words from other words, and we show that a small number of elements in word2vec vectors are sufficient to create an effective decision tree. We also show that there is "noise" in the vector space that is irrelevant to making the distinction.

The rest of this paper is organized as follows. In section 2, we introduce the background and related work. In section 3, we discuss our previous work, which serves as a motivation for this study, and our strategy in this study. We introduce our hypotheses and their verification strategy in section 4. Section 5 describes the experimental procedure. We analyze the results of the experiment in section 6. Section 7 concludes this paper and outlines future work.

2 Background and Related Work

2.1 WordNet and the Japanese WordNet

2.1.1 Princeton WordNet

Princeton WordNet [7] is a large English lexical database. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms, called a "Synset," with each expressing a distinct concept. Synsets are interlinked by conceptual-semantic and lexical token relations. One of the common relations is the hypernym–hyponym relation. WordNet can be used as a thesaurus because words are grouped in Synsets by their sense. A word can be a member of two or more Synsets because a word can be polysemous.

2.1.2The Japanese WordNet

The Japanese WordNet is a Japanese language lexical database based on Princeton WordNet. It has been in development since 2006. The structure of the Japanese WordNet was inherited from Princeton WordNet (Fig. 1). Because of differences between the Japanese and English languages, the Japanese WordNet contains original Synsets (concepts) [1]. In the Japanese WordNet, more emphasis is placed on comprehensiveness rather than accuracy [2]. It is officially stated [2] that the current version of the Japanese WordNet contains errors in approximately 5% of the entries.

2.2 Erroneous synonyms in the Japanese WordNet

Hirao et al. [3] proposed two methods for detecting erroneous synonyms in the Japanese WordNet. Their methods relied on linked structures within the Japanese WordNet and did not use corpora or text examples.

2.3 Word2vec

Word2vec [5, 6] is a tool that generates a vector representation of a word from a large amount of text data. Several studies show that word2vec is effective in finding related words by calculating the CS between words. Mikolov et al. reported an accuracy of 55.3% [6] for a rather complex task. Yamada et al. [8] reported that hyponym relations can be detected from the CSs of vectorized words.

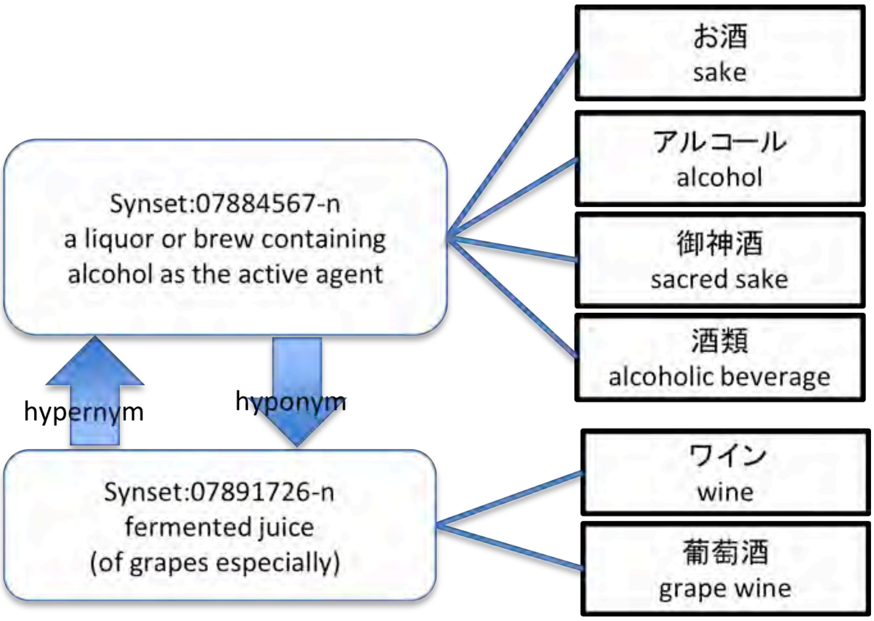

Figure 1: Synset–synonym links in the Japanese WordNet; rounded squares on the left are Synsets. Each square on the right shows a word. Lines indicate Synset–synonym pairs. Arrows indicate hypernym–hyponym relation. Only a part of the link structure is shown.

3 Previous Work

We assumed that word2vec is effective in detecting erroneous synonyms in the Japanese WordNet. First, we generated word2vec vectors for approximately 7,000 nouns in the Japanese WordNet. We used full articles from Wikipedia [9] written in the Japanese as the corpus. Next, we took the CS of each synonym pair in each Synset. Then, Synsets were sorted in ascending order of minimum CS in each Synset S (MCS(S)). For example, Synset 05752544-n comprises three words whose approximate meanings in English are {study, learning, learning-by-discipline}. The CSs of Japanese word pairs in the Synset are CS(study, learning) = 0.281, CS(study, learning-by-discipline) = 0.205, CS(learning, learning-by-discipline) = 0.516. Then, the minimum CS through all pairs of words in the Synset is MCS(105752544-n) = 0.205.

Finally, for each rank of MCS, the precision, recall, and F-value were calculated for {X|MCS(X) < MCS(A), where X ∈Synset, A is the n^th rank Synset in the MCS order}. We used 161 Synset errors used in [4] as the gold standard.

Fig. 2 shows the results. A linear increase in recall indicates that the error rate does not change with MCS. Along with the low precisions and F-values, the results indicate that CS is not a good measure of erroneous synonyms in the Japanese WordNet.

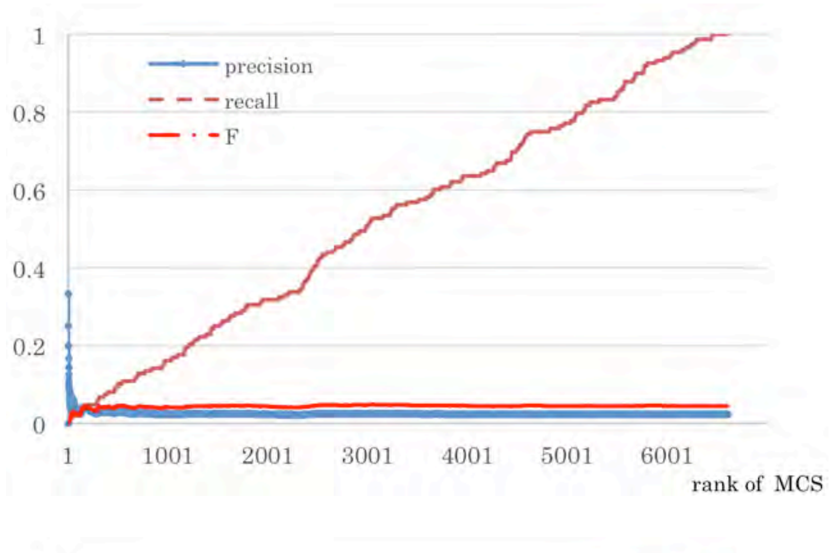

Figure 2: Result of erroneous synonym detection in the Japanese WordNet using word2vec.

4 Hypotheses on the Cause of Failure

We posed the following hypotheses on the above results:

- Word2vec vectors do indicate the sense of synonyms in a Synset.

- There is "noise" in the vector space that diminishes the usefulness of CS values for the Japanese WordNet synonyms. Only a subspace of the vector space is relevant in deciding whether a word in the Japanese WordNet is synonymous with other words.

4.1 Verification strategy

A typical Synset in the Japanese WordNet contains one to eight synonyms. It is difficult to verify the above hypotheses from such small samples. Aggregating results from randomly selected Synsets might not work because the relevant subspace can vary from one Synset to another.

We decided to use words in hyponyms of a Synset as the related words, instead of the synonyms in the Synset themselves. We assume that the set of hyponyms has a "collective sense" in the word2vec vector space.

The reason we excluded the words in the Synset themselves was to eliminate the possible influence of erroneous synonyms in the target Synsets.

5 Experiment

5.1 Generation of a word vector using word2vec

We generate word vectors of words in the Japanese WordNet using word2vec (cbow). Full Japanese articles from Wikipedia were used as the corpus, which was morphologically analyzed using MeCab [10]. Two vector sets of 200 and 800 dimensions were generated.

5.2 Target Synsets and related words

We selected 10 Synsets each of which has a link (such as hypernym, hyponym, sibling, cousin, etc.) to Synset 07891726-n as the target Synsets. They also satisfied the following conditions:

- Each target Synset has hyponym Synset
- All synonyms in each hyponym Synset appear in the corpus
- If a hyponym Synset has a hyponym of the hyponym Synsets, then all of the synonyms in the hyponym of the hyponym Synset appear in the corpus.

Table 1: Target Synsets and related words

Synset-id	related words	w.c.*
00020090-n	卵黄, 食物, 栄養分,… 'egg yolk', food, nutriment	63
00021265-n	卵黄, 飯, 食事,… 'egg yolk', 'boiled rice,' meal	108
07566340-n	米飯, 米, 飯米, … 'boiled rice', rice, cereal	107
07802417-n	穀物, 小麦, 豆, 米麦, … grain, wheat, bean, ‘food grain’	36
07881800-n	牛乳, 生乳, 粉乳, … milk, 'raw milk', 'milk powder'	69
07884567-n	ビール, クワス, 醸造 … beer, kvass, brew	34
07891726-n	赤ワイン, モーゼル, シャンパン, … ‘red wine’ , Moselle, champagne	10
14707820-n	メチルアルコール, 木精, 酒精,… methanol, 'wood alcohol' , alcohol	9
14875077-n	木炭, コークス, 軽油 ,… charcoal, coke, ‘diesel oil’	21
14940386-n	チョコレート, ココア, 果汁,… chocolate, cocoa, 'fruit juice'	51

*w.c. (word count) is the number of related words.

We used all synonyms in the hyponym-of-hyponym Synsets as well as all synonyms in the hyponym Synsets as the set of related words for the target Synset.

We manually identified that there were two Synsets (Synset 00021265-n and Synset 07566340-n) each of which contained an erroneous word in its related word list due to the incompleteness of the Japanese WordNet. We didn’t exclude those erroneous words in the following experiment. The list of full target Synset-ids and a portion of the list of related words for each target Synset are listed in Table 1.

5.3 Construction of decision trees

We built a binary decision tree that identifies related words in each target Synset. Each element in the word2vec vector is a candidate for being an input variable. As the learning set, we used all related words of a target Synset as positive examples.

Negative examples were selected using the following process. First, a human operator pick up a word randomly from the corpus which is apparently not related to the target Synset. Next, we check that the word does not have any links to the target Synset in the Japanese WordNet. Then, we append each synonym of the selected word to the negative examples if the synonym is in the corpus. After that, we repeat the process until the number of negative examples reaches that of the positive examples.

We applied the classification and regression trees (CART) algorithm [11], which uses the Gini coefficient as the partition method. The maximum depth of the trees was restricted to three. As shown in section 6, majority of the classification results are good even with the restriction. Two kinds of decision trees were constructed for 200- and 800-dimensional vectors.

5.3.1 Example of tree construction

We explain the procedure for Synset 0781726-n as an example. We already have 800- or 200-dimensional word2vec vectors. Hyponym and hyponym-of-hyponym Synsets of 0781726-n contain synonyms each of which has the following approximate meaning in English: {red-wine, Moselle, champagne, champagne (another word that means "champagne" in Japanese), Bordeaux, vermouth, port-wine, sherry, Muscat, bishop-port-wine}. All the Japanese synonyms are in the Japanese Wikipedia corpus. We selected those words as the positive examples in the learning set. The negative examples were constructed as explained in section 5.2.

The CART algorithm was used to construct the binary decision tree shown in Fig. 3. "X[706] <= -0.0368" in the node at the top of the decision tree indicates that the 706th element of word2vec vectors is used as the deciding variable in the root node.

At the bottom of each node, "samples = n" indicates the number of positive examples + negative examples that reached the node. In the three leaf nodes, "value = [e, d]" indicates the number of positive and negative examples that reached the leaf nodes. In this example, every leaf node contains only positive or negative examples. All words in the positive and negative examples are classified correctly by the decision tree. The Gini coefficient used in the algorithm is indicated by "gini=r."

6 Result

6.1 Usefulness of word2vec

Table 2 shows the results of the experiment. Precision, recall, and F-measure are calculated as follows.

First, we select all the leaf nodes that contain positive example selection results. Only one of the two leaves with a higher positive example ratio is the result of a positive example selection when a node has two leaves. When a node has a leaf and a child inner node, the leaf is the result of positive example selection only if it has a higher positive example ratio than that of the child inner node. In Fig. 3, leaves (1) and (3) are the results of positive example selection.

The number of true positive examples is the total of the positive examples in those leaves, and similarly, the number of false positives is the total of the negative examples in the same leaves.

True and false negatives are calculated from the aggregation of the results in the remaining leaves. In Fig. 3, only node (2) is the remaining leaf.

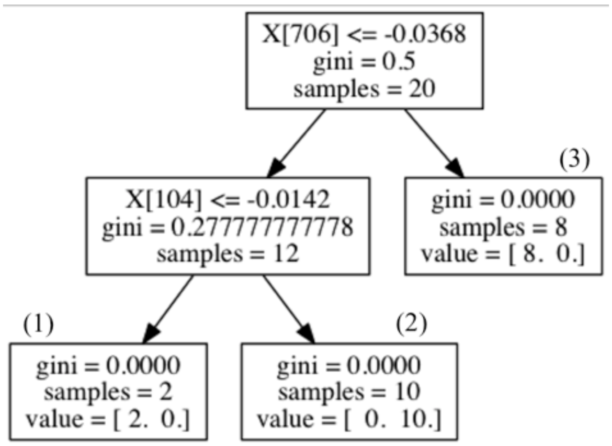

Figure 3: Decision tree for target Synset 07891726-n in the Japanese WordNet for the 800-dimensional word2vec vector.

Although the results in Table 2 are not taken from independent test data, the scores are remarkably high when the restriction of max depth <=3 is taken into account. Particularly, for 14707828-n, one vector element is sufficient to distinguish 9 + 9 positive + negative examples. When we assume that vector values are random, the probability of such occurrence is p = (9! * 9! /18! * 2 * 200) = 0.004 for a 200-dimensional vector space with nine positive and nine negative examples.

The 200- and 800-dimensional vectors produce similar results except in the case of target Synset 00021265-n. The F-value and other scores of 00021265-n in the 800-dimensional vector space are exceptionally low.

We pose the following conjecture as the reason for the low quality result: *Features of a word are more precisely indicated in the 800-dimensional vector space than in the 200-dimensional space. We sometimes need more deciding conditions than are in the 200 dimensions. In this case, the limit "depth 3" is too small to distinguish 108 + 108 elements.* It should be noted that there is an erroneous word in the related word list of 00021265-n (see section 5.2).

Table 2: Classification results

Synset-id (dim. of vector)	Precision	recall	F	d.c.	w.c.
00020090-n(200)	0.9500	0.9048	0.9268	5	63
00020090-n(800)	0.9403	1.0000	0.9692	6	63
00021265-n(200)	0.9273	0.9444	0.9358	7	108
00021265-n(800)	0.7286	0.4722	0.5730	6	108
07566340-n(200)	0.9074	0.9159	0.9116	7	107
07566340-n(800)	0.9184	0.8411	0.8780	7	107
07802417-n(200)	1	1	1	5	36
07802417-n(800)	1	0.9444	0.9714	5	36
07881800-n(200)	0.9200	1	0.9583	6	69
07881800-n(800)	0.9692	0.9130	0.9403	7	69
07884567-n(200)	1	1	1	5	34
07884567-n(800)	0.9714	1	0.9855	5	34
07891726-n(200)	1	1	1	2	10
07891726-n(800)	1	1	1	2	10
14708720-n(200)	1	1	1	1	9
14708720-n(800)	1	1	1	1	9
14875077-n(200)	1	1	1	3	21
14875077-n(800)	1	1	1	3	21
14940386-n(200)	0.9273	1	0.9623	6	51
14940386-n(800)	0.9808	1	0.9903	6	51

*d.c. (decision node count) is the number of vector elements used in a decision tree. w.c. is the number of related words, as in Table 1.

The overall results suggest that word2vec could be useful for identifying related words in the Japanese WordNet. Our first hypothesis is highly plausible.

6.2 Noise in vector space

Now, we will investigate the second hypothesis regarding "noise" in vectors. The results in Table 2 indicate that fewer than eight elements of vectors can distinguish related words of a target Synset. We must examine vector elements that are not used in the decision tree. We calculated the coefficients of variation (C.V.) for each element of the vectors in positive and negative examples.

|C.V.|(X[i]) = | StdDev(X[i]) /Average (X[i])| (X[i] is a vector element, i = 0 to 799).

Fig. 4 is the histogram of |C.V.| for all elements in vectors for target Synset 07891726-n with 10 positive and negative example words in 800-dimensional space. The histogram of |C.V.| for 10 randomly generated nouns is shown in Fig. 4 as a comparison.

In the positive examples, the ratio of the elements for which 0 ≦ |C.V.| < 3.0 is greater than the corresponding ratio in the negative examples. Features of related words can be expressed in those vector elements (0 ≦ |C.V.| < 3.0). For two vector elements used in the decision tree, X[104] and X[706] (see Fig. 3), the |C.V.| values of X[104] and X[706] are 0.773 and 0.680, respectively.

In the positive examples of 07891726-n, more than 46% of the vector elements have |C.V.| ≧ 3.0. An element with |C.V.| ≧ 3.0 can take +/- values almost randomly throughout the positive and negative examples. This could be a source of noise in the calculation of the CS of the related words.

In all 10 decision trees for the 800-dimensional vectors listed in Table 2, the |C.V.| of all vector elements used in the "root decision node" are less than 1.4. The |C.V.| is less than 3.0 in all 17 vector elements used in the second-level decision nodes. However, on the third level nodes, there are five vector elements for which |C.V.| is greater than 3.0.

One possible reason that vector elements on the third level have a greater |C.V.| is the following: *a set of words used in a lower-level decision node can have features different from those of the original set of related words after being filtered through the higher decision nodes.* For example, related words of Synset 00020090-n (substance) contain both "poison" and "food" (see section 5.2 for the construction of related words). |C.V.| values of some vector elements in the original related words are meaningless when "poison" is classified into a different subset from that of "food."

Fig. 4 shows a greater number of low |C.V.| (less than 3.0) vector elements in negative examples of 07891726-n than in the randomly selected nouns. Our negative example construction method tends to select two or more synonyms from the same Synset (see section 5.3). Therefore it is understandable that a set of vectors containing multiple synonyms in it indicates lower |C.V.| values in some vector elements. Choice of the construction method may have affected the performance of the decision trees discussed in section 6.1.

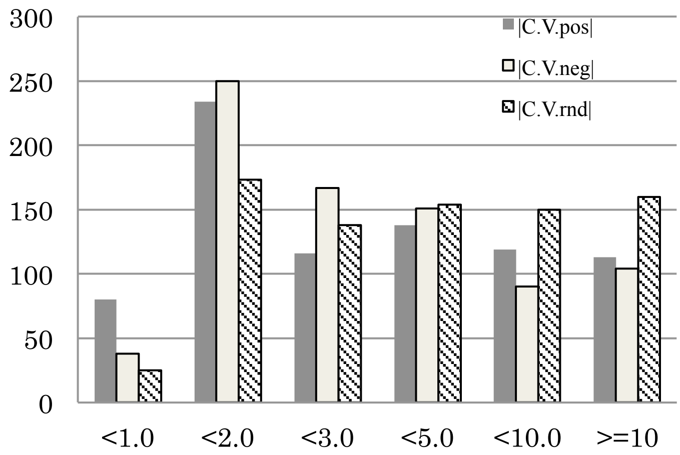

Figure 4: Histogram of |C.V.| for target Synset 07891726-n in 800-dimensional vectors. |C.V.pos| is for positive examples, |C.V.neg| is for negative examples, and |C.V.rnd| is calculated from randomly selected noun.

6.2.1 Noise in a Synset

442 Synsets (6.7%) in the 7,000 Synsets have less than zero MCS. We randomly picked 10 error-free noun Synsets, each of which has no fewer than 10 synonyms. Then, we calculated the |C.V.| between synonyms in each Synset; 30% to 58% of vector elements have |C.V.| ≧ 3.0 (800-dimensional vectors). There are two Synsets that have negative minimum CS in the 10 Synsets.

The overall results support the hypothesis that there are "noise" elements in the word2vec vector space that obstruct CS identification of erroneous synonyms in the Japanese WordNet.

7 Future Work

In this paper, we showed evidence of the following:

-A small subspace of word2vec vectors is sufficient to identify related words in the Japanese WordNet.

-There is "noise" in the vector spaces that diminishes the usefulness of CS values in distinguishing related words.

Our next step is the detection of erroneous synonyms in a single Synset of the Japanese WordNet by utilizing the abovementioned properties of word2vec and the Japanese WordNet. Although some of the vector elements with low |C.V.| are used in decision trees, there remain many vector elements with low |C.V.| that are not yet used. Random forests (RF) [12] built from low-|C.V.| vector elements are expected to be useful. We have a plan to check all erroneous synonyms in noun Synsets in the Japanese WordNet by using RF.

References

[1] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, Enhancing the Japanese WordNet, ALR7 Proc. the 7th Workshop on Asian Language Resources, pp. 1-8 ,Association for Computational Linguistics. pp. 1-8, 2009

[2] http://compling.hss.ntu.edu.sg/wnja/index.en.html

[3] T. Hirao, T. Suzuki, K. Miyata, S. Hirokawa, Detection Methods for Misplacement of Synonyms in the Japanese WordNet, International Journal of Computer & Information Science, vol. 15, no.2, pp.26-35, 2014.

[4] T. Hirao, T. Suzuki, K. Miyata, S. Hirokawa, A Trial for Detecting Misplacement of Synonyms in the Japanese WordNet using Corpus (in Japanese), ICIEC Technical report, vol. 114, no. 339, AI2014-18, pp. 13-18, 2014.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[6] T. Mikolov et al. ,Distributed representations of words and phrases and their compositionality Proc. 27th Annual Conference on Neural Information Processing Systems,

2013.

[7] Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu.

[8] I. Yamada et al, Construction of the Set of Instances from Hypernym-Hyponym relations (In Japanese). ICIEC Technical report, NLC2014-55, 2015.

[9] http://ja.wikipedia.org/wiki/.

[10] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (MeCab).

[11] L. Breiman, J. Friedman, C. J. Stone, R.A. Olshen, Classification and Regression Trees, Wadsworth & Brooks, 1984.

[12] L. Breiman, Random Forests, Machine Learning, Volume 45, Issue 1, pp 5-32, 2001.