Evaluation of Word Vector Representations by Subspace Alignment

Yulia Tsvetkov Manaal Faruqui Wang Ling Guillaume Lample Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{ytsvetko, mfaruqui, lingwang, glample, cdyer}@cs.cmu.edu

Abstract

Unsupervisedly learned word vectors have proven to provide exceptionally effective features in many NLP tasks. Most common intrinsic evaluations of vector quality measure correlation with similarity judgments. However, these often correlate poorly with how well the learned representations perform as features in downstream evaluation tasks. We present QVEC—a computationally inexpensive intrinsic evaluation measure of the quality of word embeddings based on alignment to a matrix of features extracted from manually crafted lexical resources—that obtains strong correlation with performance of the vectors in a battery of downstream semantic evaluation tasks.[1]

1 Introduction

A major attraction of vector space word representations is that they can be derived from large unannotated corpora, and they are useful as a source of features for downstream NLP tasks that are learned from small amounts of supervision. Unsupervised word vectors have been shown to benefit parsing (Lazaridou et al., 2013; Bansal et al., 2014), chunking (Turian et al., 2010), named entity recognition (Guo et al., 2014) and sentiment analysis (Socher et al., 2013), among others.

Despite their ubiquity, there is no standard scheme for intrinsically evaluating the quality of word vectors: a vector quality is traditionally judged by its utility in downstream NLP tasks. This lack of standardized evaluation is due, in part, to word vectors' major criticism: word vectors are linguistically opaque in a sense that it is still not clear how to interpret individual vector dimensions, and, consequently, it is not clear how to score a non-interpretable representation. Nevertheless, to facilitate development of better word vector models and for better error analysis of word vectors, it is desirable (1) to compare word vector models easily, without recourse to multiple extrinsic applications whose implementation and runtime can be costly; and (2) to understand how features in word vectors contribute to downstream tasks.

We propose a simple intrinsic evaluation measure for word vectors. Our measure is based on component-wise correlations with manually constructed "linguistic" word vectors whose components have well-defined linguistic properties (§2). Since vectors are typically used to provide features to downstream learning problems, our measure favors *recall* (rather than precision), which captures our intuition that meaningless dimensions in induced vector representations are less harmful than important dimensions that are missing. We thus align dimensions in a distributional word vector model with the linguistic dimension vectors to maximize the cumulative correlation of the aligned dimensions (§3). The resulting sum of correlations of the aligned dimensions is our evaluation score. Since the dimensions in the linguistic vectors are linguistically-informed, the alignment provides an "annotation" of components of the word vector space being evaluated.

To show that our proposed score is meaningful, we compare our intrinsic evaluation model to the standard (semantic) extrinsic evaluation benchmarks (§4). For nine off-the-shelf word vector representation models, our model obtains high correlation ($0.34 \leq r \leq 0.89$) with the extrinsic tasks (§5).

2 Linguistic Dimension Word Vectors

The crux of our evaluation method lies in quantifying the similarity between a distributional word vector model and a (gold-standard) linguistic re-

[1] The evaluation script and linguistic vectors described in this paper are available at
https://github.com/ytsvetko/qvec

source capturing human knowledge. To evaluate the semantic content of word vectors, we exploit an existing semantic resource—SemCor (Miller et al., 1993). From the SemCor annotations we construct a set of linguistic word vectors, details are given in the rest of this section; table 1 shows an example of the vectors.

WordNet (Fellbaum, 1998, WN) partitions nouns and verbs into coarse semantic categories known as supersenses (Ciaramita and Altun, 2006; Nastase, 2008).[2] There are 41 supersense types: 26 for nouns and 15 for verbs, for example, NOUN.BODY, NOUN.ANIMAL, VERB.CONSUMPTION, or VERB.MOTION. SemCor is a WordNet-annotated corpus that captures, among others, supersense annotations of WordNet's 13,174 noun lemmas and 5,686 verb lemmas at least once. We construct term frequency vectors normalized to probabilities for all nouns and verbs that occur in SemCor at least 5 times. The resulting set of 4,199 linguistic word vectors has 41 interpretable columns.

WORD	NN.ANIMAL	NN.FOOD	···	VB.MOTION
fish	0.68	0.16	···	0.00
duck	0.31	0.00	···	0.69
chicken	0.33	0.67	···	0.00

Table 1: Oracle linguistic word vectors, constructed from a linguistic resource containing semantic annotations.

3 Word Vector Evaluation Model

We align dimensions of distributional word vectors to dimensions (linguistic properties) in the linguistic vectors described in §2 to maximize the cumulative correlation of the aligned dimensions. By projecting linguistic annotations via the alignments, we also obtain plausible annotations of dimensions in the distributional word vectors. In this section, we formally describe the model, which we call the QVEC.

Let the number of common words in the vocabulary of the distributional and linguistic word vectors be N. We define, the distributional vector matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$ with every row as a dimension vector $\mathbf{x} \in \mathbb{R}^{1\times N}$. D denotes word vector dimensionality. Similarly, $\mathbf{S} \in \mathbb{R}^{P\times N}$ is the linguistic property matrix with every row as a linguistic property vector $\mathbf{s} \in \mathbb{R}^{1\times N}$. P denotes linguistic properties obtained from a manually-annotated linguistic resource. We obtain an alignment between the word vector dimensions and the linguistic dimensions which maximizes the correlation between the aligned dimensions of the two matrices. This is 1:n alignment: one distributional dimension is aligned to at most one linguistic property, whereas one linguistic property can be aligned to n distributional dimensions; see figure 1.

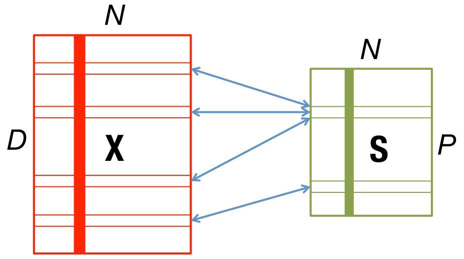

Figure 1: The filled vertical vectors represent the word vector in the word vector matrix $\mathbf{X}$ and the linguistic property matrix $\mathbf{S}$. The horizontal hollow vectors represent the "distributional dimension vector" in $\mathbf{X}$ and "linguistic dimension vector" in $\mathbf{S}$. The arrows show mapping between distributional and linguistic vector dimensions.

Let $\mathbf{A} \in \{0,1\}^{D\times P}$ be a matrix of alignments such that $a_{ij} = 1$ iff $\mathbf{x}_i$ is aligned to $\mathbf{s}_j$, otherwise $a_{ij} = 0$. If $r(\mathbf{x}_i, \mathbf{s}_j)$ is the Pearson's correlation between vectors $\mathbf{x}_i$ and $\mathbf{s}_j$, then our objective is defined as:

$$\text{QVEC} = \max_{\mathbf{A}|\sum_j a_{ij}\leq 1} \sum_{i=1}^{D}\sum_{j=1}^{P} r(\mathbf{x}_i, \mathbf{s}_j) \times a_{ij} \quad (1)$$

The constraint $\sum_j a_{ij} \leq 1$, warrants that one distributional dimension is aligned to at most one linguistic dimension. The total correlation between two matrices QVEC is our intrinsic evaluation measure of a set of word vectors relative to a set of linguistic properties.

The QVEC's underlying hypothesis is that dimensions in distributional vectors correspond to linguistic properties of words. It is motivated, among others, by the effectiveness of word vectors in linear models implying that linear combinations of features (vector dimensions) produce relevant, salient content. Via the alignments a_{ij} we obtain labels on dimensions in the distributional word vectors. The magnitude of the correlation $r(\mathbf{x}_i, \mathbf{s}_j)$ corresponds to the annotation confidence: the higher the correlation, the more salient the linguistic content of the dimension. Clearly, dimensions in the linguistic matrix S do not capture every possible linguistic property, and low correlations often correspond to the missing information in the linguistic matrix. Thus, QVEC is a recall-oriented measure: highly-

[2] Supersenses are known as "lexicographer classes" in WordNet documentation, `http://wordnet.princeton.edu/man/lexnames.5WN.html`

correlated alignments provide evaluation and annotation of vector dimensions, and missing information or noisy dimensions do not significantly affect the score since the correlations are low.

4 Experimental Setup

4.1 Word Vector Models

To test the QVEC, we select a diverse suite of popular/state-of-the-art word vector models. All vectors are trained on 1 billion tokens (213,093 types) of English Wikipedia corpus with vector dimensionality 50, 100, 200, 300, 500, 1000.

CBOW and Skip-Gram (SG). The WORD2VEC tool (Mikolov et al., 2013) is fast and widely-used. In the SG model, each word's Huffman code is used as an input to a log-linear classifier with a continuous projection layer and words within a given context window are predicted. In the CBOW model a word is predicted given the context words.[3]

CWindow and Structured Skip-Gram (SSG). Ling et al. (2015b) propose a syntactic modification to the WORD2VEC models that accounts for word order information, obtaining state-of-the-art performance in syntactic downstream tasks.[4]

CBOW with Attention (Attention). Ling et al. (2015a) further improve the WORD2VEC CBOW model by employing an attention model which finds, within the contextual words, the words that are relevant for each prediction. These vectors have been shown to benefit both semantically and syntactically oriented tasks.

GloVe. Global vectors for word representations (Pennington et al., 2014) are trained on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations show interesting linear substructures of the vector space.[5]

Latent Semantic Analysis (LSA). We construct word-word co-occurrence matrix $\mathbf{X}$; every element in the matrix is the pointwise mutual information between the two words (Church and Hanks, 1990). Then, truncated singular value decomposition is applied to factorize $\mathbf{X}$, where we keep the k largest singular values. Low dimensional word vectors of dimension k are obtained from $\mathbf{U_k}$ where $\mathbf{X} \approx \mathbf{U_k}\Sigma\mathbf{V_k}^\top$ (Landauer and Dumais, 1997).

GloVe+WN, GloVe+PPDB, LSA+WN, LSA+PPDB. We use retrofitting (Faruqui et al., 2015) as a post-processing step to enrich GloVe and LSA vectors with semantic information from WordNet and Paraphrase database (PPDB) (Ganitkevitch et al., 2013).[6]

4.2 Semantic Evaluation Benchmarks

We compare the QVEC to six standard extrinsic semantic tasks for evaluating word vectors; we now briefly describe the tasks.

Word Similarity. We use three different benchmarks to measure word similarity. The first one is the **WS-353** dataset (Finkelstein et al., 2001), which contains 353 pairs of English words that have been assigned similarity ratings by humans. The second is the **MEN** dataset (Bruni et al., 2012) of 3,000 words pairs sampled from words that occur at least 700 times in a large web corpus. The third dataset is **SimLex-999** (Hill et al., 2014) which has been constructed to overcome the shortcomings of WS-353 and contains 999 pairs of adjectives, nouns and verbs. Word similarity is computed using cosine similarity between two words and the performance of word vectors is computed by Spearman's rank correlation between the rankings produced by vector model against the human rankings.[7]

Text Classification. We consider four binary categorization tasks from the 20 Newsgroups **(20NG)** dataset.[8] Each task involves categorizing a document according to two related categories with training/dev/test split in accordance with Yogatama and Smith (2014). For example, a classification task is between two categories of Sports: baseball vs hockey. We report the average classification accuracy across the four tasks. Our next downstream semantic task is the sentiment analysis task **(Senti)** (Socher et al., 2013) which is a binary classification task between positive and negative movie reviews using the standard training/dev/test split and report accuracy on the test set. In both cases, we use the average of the word vectors of words in a document (and sentence, respectively) and use them as features in an ℓ_2-regularized logistic regression classifier. Finally, we evaluate vectors on the metaphor detection **(Metaphor)** (Tsvetkov et al.,

[3] https://code.google.com/p/word2vec

[4] https://github.com/wlin12/wang2vec

[5] http://www-nlp.stanford.edu/projects/glove/

[6] https://github.com/mfaruqui/retrofitting

[7] We employ an implementation of a suite of word similarity tasks at wordvectors.org (Faruqui and Dyer, 2014).

[8] http://qwone.com/~jason/20Newsgroups

2014a).[9] The system uses word vectors as features in a random forest classifier to label adjective-noun pairs as literal/metaphoric. We report the system accuracy in 5-fold cross validation.

5 Results

To test the efficiency of QVEC in capturing the semantic content of word vectors, we evaluate how well QVEC's scores correspond to the scores of word vector models on semantic benchmarks. We compute the Pearson's correlation coefficient r to quantify the linear relationship between the scorings. We begin with comparison of QVEC with one extrinsic task—Senti—evaluating 300-dimensional vectors.

Model	QVEC	Senti
CBOW	40.3	90.0
SG	35.9	80.5
CWindow	28.1	76.2
SSG	40.5	81.2
Attention	40.8	80.1
GloVe	34.4	79.4
GloVe+WN	42.1	79.6
GloVe+PPDB	39.2	79.7
LSA	19.7	76.9
LSA+WN	29.4	77.5
LSA+PPDB	28.4	77.3
Correlation (r)	**0.87**	

Table 2: Intrinsic (QVEC) and extrinsic scores of the 300-dimensional vectors trained using different word vector models and evaluated on the Senti task. Pearson's correlation between the intrinsic and extrinsic scores is $r = 0.87$.

As we show in table 2, the Pearson's correlation between the intrinsic and extrinsic scores is $r = 0.87$. To account for variance in WORD2VEC representations (due to their random initialization and negative sampling strategies, the representations are different for each run of the model), and to compare QVEC to a larger set of vectors, we now train three versions of vector sets per model. This results in 21 word vector sets: three vector sets per five WORD2VEC models plus GloVe, LSA, and retrofitting vectors shown in table 2. The Pearson's correlation computed on the extended set of comparison points (in the same experimental setup as in table 2) is $r = 0.88$. In the rest of this section we report results on the extended suite of word vectors.

We now extend the table 2 results, and show correlations between the QVEC and extrinsic scores across all benchmarks for 300-dimensional vectors. Table 3 summarizes the results. The QVEC obtains high positive correlation with all the semantic tasks.

Table 4 shows, for the same 300-dimensional vectors, that QVEC's correlation with the downstream text classification tasks is on par with or higher than the correlation between the word similarity and text classification tasks. Higher correlating methods—in our experiments, QVEC and MEN—are better predictors of quality in downstream tasks.

	20NG	Metaphor	Senti
WS-353	0.55	0.25	0.46
MEN	**0.76**	0.49	0.55
SimLex	0.56	0.44	0.51
QVEC	0.74	**0.75**	**0.88**

Table 4: Pearson's correlations between word similarity/QVEC scores and the downstream text classification tasks.

Next, we measure correlations of QVEC with the extrinsic tasks across word vector models with different dimensionality. The results are shown in figure 2.

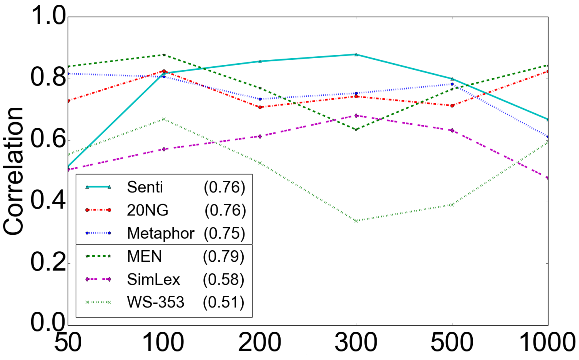

Figure 2: Pearson's correlation between QVEC scores and the semantic benchmarks across word vector models on vectors of different dimensionality. The scores at dimension 300 correspond to the results shown in table 3. The scores in the legend show average correlation across dimensions.

To summarize, we observe high positive correlation between QVEC and the downstream tasks, consistent across the tasks and across different models with vectors of different dimensionalities.

Since QVEC favors recall over precision, larger numbers of dimensions will *ceteris paribus* result in higher scores—but not necessarily higher correlations with downstream tasks. We therefore impose the restriction that QVEC only be used to compare vectors of the same size, but we now show that its correlation with downstream tasks is stable, conditional on the size of the vectors being compared. We aggregate rankings by individual

[9] https://github.com/ytsvetko/metaphor

	WS-353	**MEN**	**SimLex**	**20NG**	**Metaphor**	**Senti**
r	0.34	0.63	0.68	0.74	0.75	0.88

Table 3: Pearson's correlations between QVEC scores of the 300-dimensional vectors trained using different word vector models and the scores of the downstream tasks on the same vectors.

	50	**100**	**200**	**300**	**500**	**1000**
ρ(QVEC, Senti)	0.32	0.57	0.73	0.78	0.72	0.60
ρ(QVEC, All)	0.66	0.59	0.63	0.65	0.62	0.59

Table 5: Spearman's rank-order correlation between the QVEC ranking of the word vector models and the ranking produced by (1) the Senti task, or (2) the aggregated ranking of all tasks (All). We rank separately models of vectors of different dimensionality (table columns).

downstream tasks into a global ranking using the Kemeny–Young rank aggregation algorithm, for each dimension separately (Kemeny, 1959). The algorithm finds a ranking which minimizes pairwise disagreement of individual rankers. Table 5 shows Spearman's rank correlation between the rankings produced by the QVEC and the Senti task/the aggregated ranking. For example, ranking of 300-dimensional models produced by Senti is *{SSG, CBOW, SG, Attention, GloVe+PPDB, GloVe+WN, GloVe, LSA+WN, LSA+PPDB, LSA, CWindow}*, and the QVEC's ranking is *{GloVe+WN, Attention, SSG, CBOW, GloVe+PPDB, SG, GloVe, LSA+WN, LSA+PPDB, CWindow, LSA}*. The Spearman's ρ between the two rankings is 0.78. We note, however, that there is a considerable variation between rankings across all models and across all dimensions, for example the SimLex ranking produced for the same 300-dimensional vectors is *{GloVe+PPDB, GloVe+WN, SG, LSA+PPDB, SSG, CBOW, Attention, CWindow, LSA+WN, GloVe, LSA}*, and ρ(Senti, SimLex) $= 0.46$. In a recent related study, Schnabel et al. (2015) also observe that existing word similarity and text categorization evaluations yield different orderings of word vector models. This task-specifity of rankings emphasizes the deficiency of evaluating word vector models solely on downstream tasks, and the need of a standardized intrinsic evaluation approach that quantifies linguistic content of word vectors.

6 Future Work

Aligning dimensions of linguistic and distributional vectors enables projection of linguistic annotations via the alignments, and thereby facilitates qualitative analysis of individual dimensions in distributional vectors. Albeit noisy, we find correspondence between the projected labels of distributional columns and the column content. For example, in the 50-dimensional SG model top-10 ranked words in a dimension aligned to NOUN.BODY with r=0.26 are *amputated, sprained, palsy, semenya, lacerations, genital, cervical, concussion, congenital, abdominal*. This interesting by-product of our method will be addressed in future work.

While we experiment with linguistic vectors capturing semantic concepts, our methodology is generally applicable to other linguistic resources (Faruqui and Dyer, 2015). For example, part-of-speech annotations extracted from a treebank would yield linguistic vectors capturing syntactic content of vectors. Thus, QVEC can be used as a task-specific evaluator; we will investigate this in future work.

A useful property of supersenses (features in our linguistic vectors) is that they are stable across languages (Schneider et al., 2013; Tsvetkov et al., 2014b). Cross-lingual vector evaluation and evaluation of multilingual word vectors with QVEC is thus an additional promising research avenue.

7 Conclusion

We propose a method for intrinsic evaluation of word vectors which shows strong relationship—both linear and monotonic—with the scores/rankings produced by the downstream tasks.

Acknowledgments

We are grateful to the anonymous reviewers for constructive feedback. This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533.

References

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proc. of ACL*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proc. of ACL*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proc. of ACL (Demonstrations)*.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proc. ACL*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Noah A. Smith, and Eduard Hovy. 2015. Retrofitting word vectors to semantic lexicons. In *Proc. of NAACL*.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proc. of WWW*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proc. of NAACL*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proc. of EMNLP*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

John G. Kemeny. 1959. Mathematics without numbers. 88(4):577–591.

Thomas K Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*.

Angeliki Lazaridou, Eva Maria Vecchi, and Marco Baroni. 2013. Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proc. of EMNLP*.

Wang Ling, Lin Chu-Cheng, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015a. Not all contexts are created equal: Better word representations with variable attention. In *Proc. of EMNLP*.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015b. Two/too simple adaptations of `word2vec` for syntax problems. In *Proc. of NAACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Vivi Nastase. 2008. Unsupervised all-words word sense disambiguation with grammatical dependencies. In *Proc. of IJCNLP*, pages 7–12.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proc. of EMNLP*.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. NAACL-HLT*, pages 661–667.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014a. Metaphor detection with cross-lingual model transfer. In *Proc. ACL*, pages 248–258.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014b. Augmenting English adjective senses with supersenses. In *Proc. LREC*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*.

Dani Yogatama and Noah A Smith. 2014. Linguistic structured sparsity in text categorization. In *Proc. of ACL*.